Machine learning workflows identify a microRNA signature of insulin transcription in human tissues

Wong, Wilson W.K.; Joglekar, Mugdha; Saini, Vijit; Jiang, Guozhi; Dong, Charlotte; Chaitarvornki, Alissa; Maciag, Grzegorz; Gerace, Dario; Farr, Ryan; Satoor, Sarang; Sahu, Subhshri ; Sharangdhar, Tejaswini; Ahmed, Asma S; Chew, Yi Vee; Liuwantara, David; Heng, Benjamin; Lim, Chai K.; Hunter, Julie; Januszewski, Andrzej; Sørensen, Anja Elaine; Akil, Ammira; Gamble, Jennifer; Loudovaris, Thomas; Kay, Thomas W; Thomas, Helen; O'Connell, Philip; Guillemin, Gilles; Martin, David; Simpson, Ann; Hawthorne, Wayne; Dalgaard, Louise Torp; Ma, Ronald C; Hardikar, Anandwardhan Awadhoot

Published in:
iScience

DOI:
10.1016/j.isci.2021.102379

Publication date:
2021

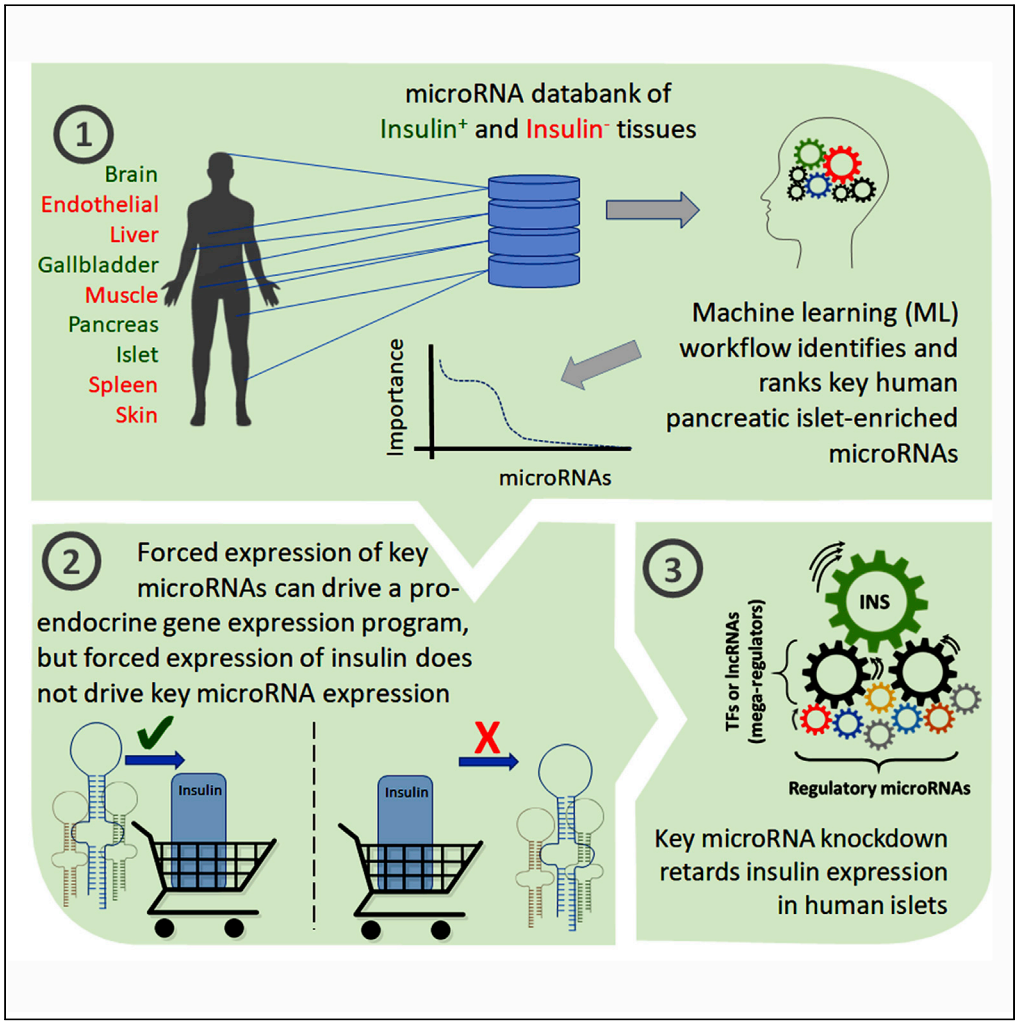
Document Version
Publisher's PDF, also known as Version of record

iScience

CellPress
OPEN ACCESS

Article

Machine learning workflows identify a microRNA signature of insulin transcription in human tissues



Wilson K.M. Wong, Mugdha V. Joglekar, Vijit Saini, ..., Louise T. Dalgaard, Ronald C.W. Ma, Anandwardhan A. Hardikar

a.hardikar@westernsydney.edu.au

Highlights

Unbiased machine learning workflow ranks miRNAs associated with insulin transcription

Forced expression of top-ranked miRNAs drives pro-endocrine program in progenitor cells

Knockdown of top-ranked miRNAs retards insulin gene transcription in human islets

Insulin transcript-associated miRNAs are reduced in islets of donors with type 2 diabetes

Wong et al., iScience 24, 102379
April 23, 2021
https://doi.org/10.1016/j.isci.2021.102379

iScience

CellPress
OPEN ACCESS

Article

Machine learning workflows identify a microRNA signature of insulin transcription in human tissues

Wilson K.M. Wong,[1,2,13] Mugdha V. Joglekar,[1,2,13] Vijit Saini,[1,3] Guozhi Jiang,[4] Charlotte X. Dong,[1,2] Alissa Chaitarvornkit,[1,2] Grzegorz J. Maciag,[5] Dario Gerace,[3] Ryan J. Farr,[1,2] Sarang N. Satoor,[1,2] Subhshri Sahu,[1,2] Tejaswini Sharangdhar,[1,2] Asma S. Ahmed,[1,2] Yi Vee Chew,[6] David Liuwantara,[6] Benjamin Heng,[7] Chai K. Lim,[7] Julie Hunter,[8] Andrzej S. Januszewski,[9] Anja E. Sørensen,[5] Ammira S.A. Akil,[10] Jennifer R. Gamble,[8] Thomas Loudovaris,[11] Thomas W. Kay,[11] Helen E. Thomas,[11] Philip J. O'Connell,[6] Gilles J. Guillemin,[7] David Martin,[12] Ann M. Simpson,[3] Wayne J. Hawthorne,[6] Louise T. Dalgaard,[5] Ronald C.W. Ma,[4] and Anandwardhan A. Hardikar[1,2,5,14,*]

SUMMARY

Dicer knockout mouse models demonstrated a key role for microRNAs in pancreatic β-cell function. Studies to identify specific microRNA(s) associated with human (pro-)endocrine gene expression are needed. We profiled microRNAs and key pancreatic genes in 353 human tissue samples. Machine learning workflows identified microRNAs associated with (pro-)insulin transcripts in a discovery set of islets (n = 30) and insulin-negative tissues (n = 62). This microRNA signature was validated in remaining 261 tissues that include nine islet samples from individuals with type 2 diabetes. Top eight microRNAs (miR-183-5p, -375-3p, 216b-5p, 183-3p, -7-5p, -217-5p, -7-2-3p, and -429-3p) were confirmed to be associated with and predictive of (pro-)insulin transcript levels. Use of doxycycline-inducible microRNA-overexpressing human pancreatic duct cell lines confirmed the regulatory roles of these microRNAs in (pro-)endocrine gene expression. Knockdown of these microRNAs in human islet cells reduced (pro-)insulin transcript abundance. Our data provide specific microRNAs to further study microRNA-mRNA interactions in regulating insulin transcription.

INTRODUCTION

Insulin production in pancreatic β-cells is an essential process required for the maintenance of normal glucose metabolism. Any decline in β-cell function is a common denominator to type 1 diabetes (T1D), as well as type 2 diabetes (T2D). Transcription of specialized genes, including (pro-)insulin, is dependent on several factors including chromatin organization, transcription factor assembly, as well as the expression of regulatory microRNAs (miRNAs/miRs). MicroRNAs are short, non-coding RNA molecules known to fine-tune gene expression via targeting multiple messenger RNAs (mRNAs). Biologically active (mature) microRNAs are generated through processing by dicer, an enzyme critical for the biogenesis of both canonical and non-canonical microRNAs (Wong et al., 2018). In mice, pancreatic *Dicer1* deletion demonstrated the essential role of microRNAs in the generation of pancreatic β-cells (Lynn et al., 2007) during embryonic development. Islet β-cell-specific (RIP-Cre) deletion of *Dicer1* in mice did not affect islet cell development but significantly reduced insulin transcription/production (Melkman-Zehavi et al., 2011) and promoted the development of diabetes (Kalis et al., 2011) in adult mice. Although these loss-of-function studies demonstrate, in general, the essential role of microRNAs in mouse islet biology, it is important to identify microRNAs associated with human (pro-)endocrine gene expression and more specifically those associated with (pro-) insulin gene transcription.

One of the very first studies in microRNA islet biology was the demonstration of a key role for miR-375 in mouse insulin secretion (Poy et al., 2004), followed by validation of its role in zebrafish (Kloosterman et al., 2007) and mouse (Poy et al., 2009) pancreatic islet development. These studies laid the foundation for

[1]Diabetes and Islet Biology Group, School of Medicine, Western Sydney University, Narellan Road & Gilchrist Drive, Campbelltown, NSW 2560, Australia

[2]Diabetes and Islet Biology group, Faculty of Medicine and Health, University of Sydney, 92-94 Parramatta Road, Camperdown, NSW 2050, Australia

[3]School of Life Sciences and the Centre for Health Technologies, University of Technology Sydney, 15 Broadway, Ultimo, NSW 2007, Australia

[4]Department of Medicine and Therapeutics, and Hong Kong Institute of Diabetes and Obesity, and Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Prince of Wales Hospital, Hong Kong, Special Administrative Region, China

[5]Department of Science and Environment, Roskilde University, Universitetsvej 1, 4000 Roskilde, Denmark

[6]Centre for Transplant and Renal Research, Westmead Institute for Medical Research, University of Sydney, 176 Hawkesbury Road, Westmead, NSW 2145, Australia

[7]Faculty of Medicine Health and Human Sciences, Macquarie University, Sydney, NSW 2019, Australia

[8]Centre for the Endothelium, Vascular Biology Program, Centenary Institute, University of Sydney Medical School, Locked Bag #6,

Continued

identification of miR-375 as a microRNA associated with and necessary for islet beta-cell function in zebrafish and mice. In human studies, miR-375 expression was shown in the developing and adult pancreatic islet cells (Klein et al., 2013; Joglekar et al., 2009a, 2009c; Correa-Medina et al., 2009) but also in islet non-β cells (Joglekar et al., 2009a) as well as in multiple non-pancreatic human tissues (Latreille et al., 2015). These studies confirmed that miR-375, although enriched in human islets, is not an "islet-specific" microRNA. It became apparent that a set or combination of microRNAs, rather than a single microRNA, supports/promotes endocrine pancreatic cell function. The scarcity of human islets for research and the availability of elegant molecular tools in non-human model systems have resulted in a large body of islet microRNA research in zebrafish and mice. Studies comparing microRNA profiles of cells/tissues that naturally transcribe the (pro-)insulin gene (e.g. thymus, gallbladder, brain (Choi et al., 2019; Devaskar et al., 1994; Heller et al., 2010; Mehran et al., 2012; Sahu et al., 2009; Dutton et al., 2007; Joglekar et al., 2021) are also lacking. Here, we sought to identify microRNAs that fine-tune the expression of the human (pro-)insulin gene using an unbiased, machine learning discovery approach, followed by wet-lab validation. Our data present a signature of microRNAs that is not only associated with (pro-)insulin transcript abundance and predictive of (pro-)insulin gene expression but also regulates the expression of pancreatic transcription factors and islet (pro-)hormones.

RESULTS

Study samples

To generate a signature of microRNAs associated with (pro-)insulin gene transcription, we isolated RNA from a total of 353 human samples across nine tissue types (Figure 1A). We then assessed (pro-)insulin as well as the housekeeping 18S rRNA gene transcripts in these samples and classified these 353 tissue samples as insulin-positive or insulin-negative tissues (Table S1). A subset (discovery set; n = 92 of 353) of these samples were then subjected to microRNA discovery analyses using the OpenArray high-throughput TaqMan-based real-time quantitative PCR (qPCR) platform (Farr et al., 2015; Wong et al., 2015). The discovery set consisted of two groups of tissues that are at the ends of the insulin transcript expression spectrum: one with very high insulin expression (pure islets) and another with undetectable insulin transcripts. The selection of samples in each group was random. Along with the 754 different microRNAs, we also assayed candidate mRNAs for islet hormones (*INS*, *GCG*, *SST*), transcription factors (*PDX1*, *MAFA*, *NEUROG3*; "referred herein as" *NGN3*, *HES1*), and the housekeeping gene (18S rRNA; Figure S1A) in this discovery set of 92 samples that include 30 human islets and 62 insulin-negative samples (Table S1). The remaining samples were then used in validation (n = 50) and prediction (n = 202) sets of human tissues (Table S1). A separate set of nine human organ donor islets from individuals with T2D were also used to assess microRNA expression.

Identifying microRNAs associated with human pancreatic islet (pro-)insulin expression

We performed unsupervised bidirectional hierarchical clustering on global normalized microRNA as well as candidate mRNA expression data from our discovery sample set (n = 92). All human islet samples (insulin-positive tissues) cluster together indicating similarity in their microRNA (Figure 1B) as well as candidate mRNA expression profiles (Figure S1A), relative to those from insulin-negative tissues. Interestingly, a total of 241 different microRNAs were found at significantly higher levels (≥2-fold, $p < 0.05$) in human islets, and 91 other microRNAs were significantly higher in insulin-negative tissues (≥2-fold, $p < 0.05$; Figure 1C and Table S2). All of these 332 differentially expressed microRNAs were expressed between 2- to over a hundred thousand-fold higher abundance (Cycle threshold [Ct] value difference of 1–17; Figure 1C and Table S2) and at significantly low p value ($p < 0.05$ to $p < 3.4 \times 10^{-49}$), relative to the insulin-negative tissues and remain significant after adjustment for multiple testing. Several of these microRNAs show a high correlation with the candidate mRNAs assessed, as presented by connecting lines in the center of the Circos plot (correlation coefficient (r) ≥0.6; Figure 1D).

Since conventional comparisons of microRNA expression data sets (Figures 1B–1D) identified a large number of significantly differentially expressed microRNAs, we decided to employ machine learning workflows to decipher microRNAs that are highly associated with (pro-)insulin (mRNA) levels. Penalization algorithms with bootstrap machine learning workflow (Figure 1E) were used to perform microRNA selection. Good quality human islets (insulin Ct value ≤16.8; n = 30) were marked as the insulin-positive (1) group and all insulin-negative tissues (insulin Ct value = 39; n = 62) were categorized under the insulin-negative (0) group. Validation of the model was carried out using bootstrapping to confirm the signature of microRNAs that were highly associated with (pro-)insulin gene expression. Bootstrap validation workflow of penalized logistic regression included resampling data for 1000 times (Figure 1E). During these analyses, ~37% of the

Newtown, NSW 2042, Australia

[9]NHMRC Clinical Trials Centre, University of Sydney, 92-94 Parramatta Road, Camperdown, NSW 2050, Australia

[10]Department of Human Genetics-Precision Medicine Program, Sidra Medicine, P.O. Box 26999, Doha, Qatar

[11]St Vincent's Institute and The University of Melbourne Department of Medicine, 9 Princes Street, Fitzroy, VIC, Australia

[12]Upper GI Surgery, Strathfield Hospital, 2/3 Everton Road, Strathfield, NSW 2135, Australia

[13]These authors contributed equally

[14]Lead contact

*Correspondence: a.hardikar@westernsydney.edu.au

https://doi.org/10.1016/j.isci.2021.102379

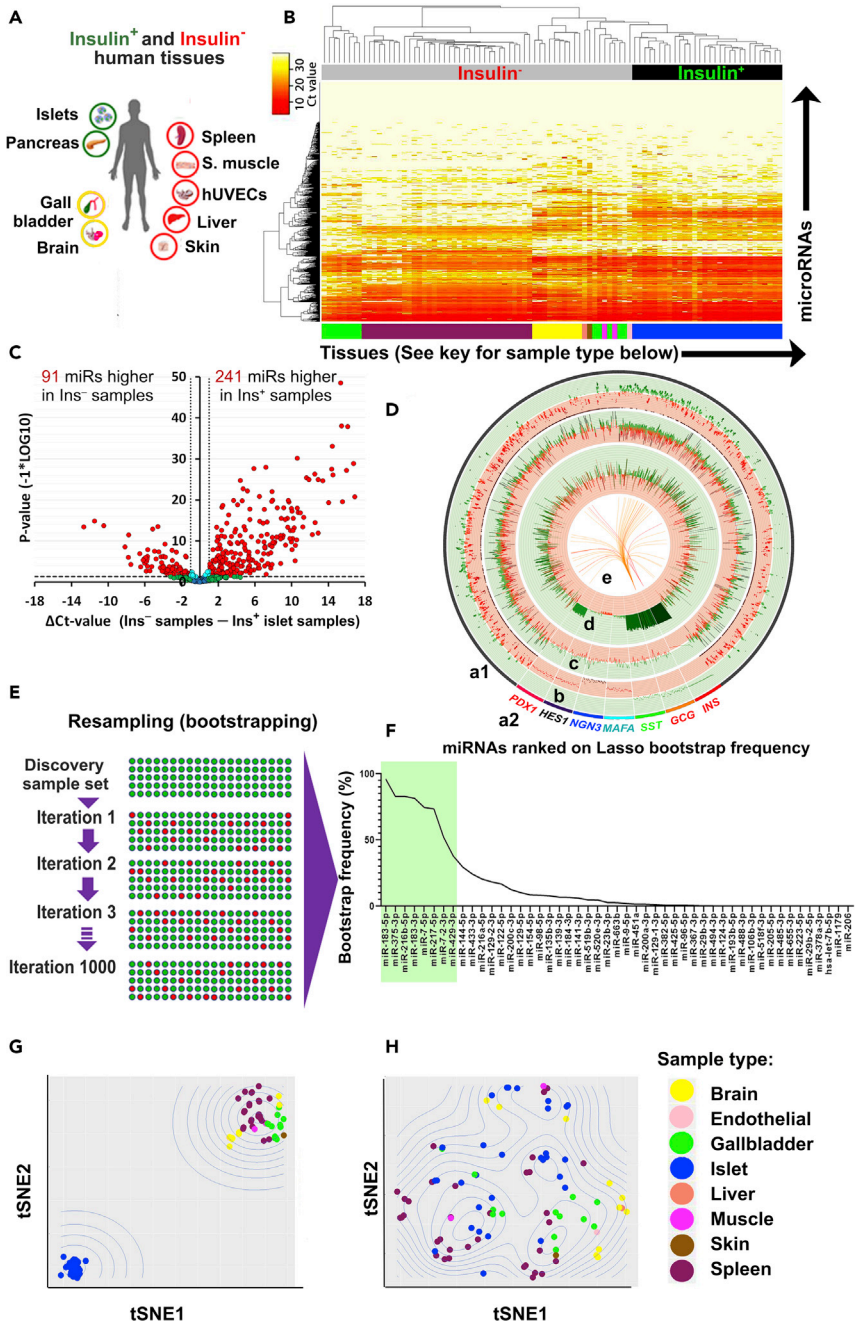


Figure 1. Identification of (pro-)insulin transcript-associated microRNAs in the discovery sample set

(A) Schematic presentation of insulin-positive (green outline) and insulin-negative (red outline) human cell/tissues used. Human cell/tissues (such as the gallbladder and brain, yellow outline) which have low to undetectable insulin expression were also used in this study.

(B) An unsupervised Euclidean (average) bidirectional hierarchical plot on the 754 discovery microRNAs in human islets (n = 30; all insulin positive) and insulin-negative tissues (n = 62). Tissues included in the discovery set are color coded, and the key to color and their corresponding tissue is provided in the bottom right corner of this figure. Heatmap represents the normalized qPCR Ct values (color bar) with low Ct values/high miRNA levels (dark red) and higher Ct values/low to no expression of microRNA (yellow to white).

(C) Volcano plot for the 754 discovery microRNAs presented as Ct value differences (on X axis) and statistical significance (−log10 p value; two-tailed Welch's t-test) on Y axis for insulin-negative vs islet (insulin positive) samples. The dashed

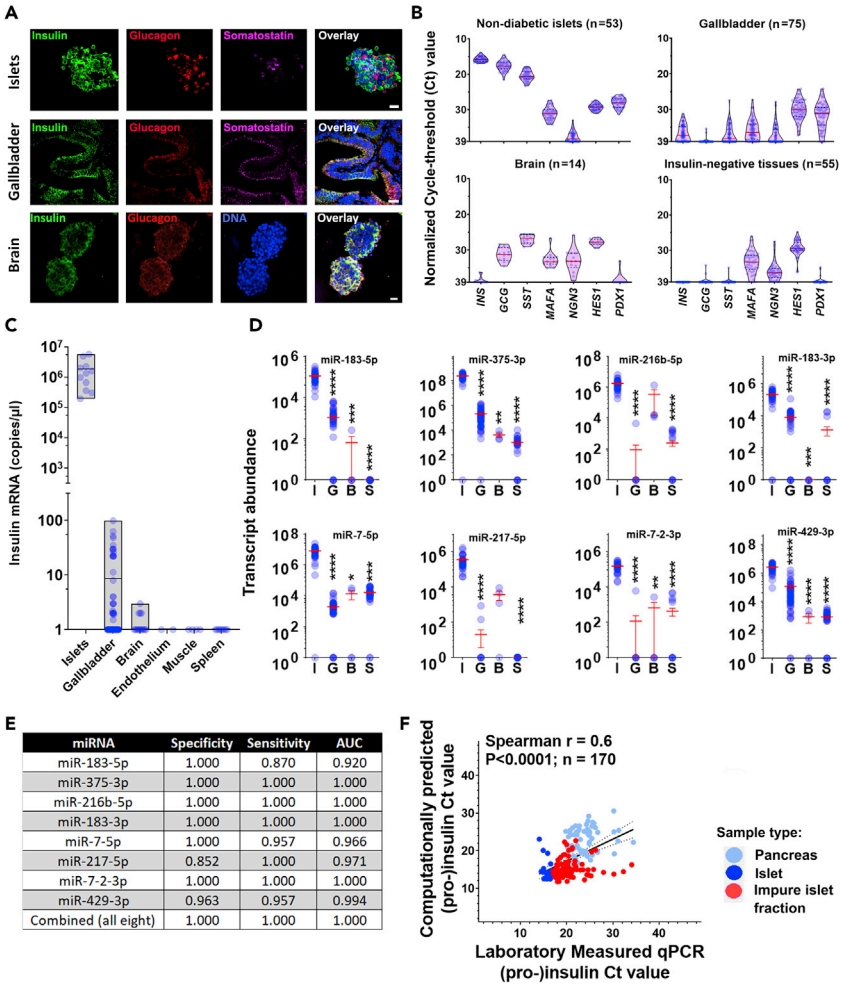
Figure 1. *Continued*
horizontal line represents p value = 0.05, whereas dotted vertical lines represent a 2-fold difference (1 Ct value). Significantly altered microRNAs are in red.
(D) A Circos plot for mRNA and microRNA expression in human islets is divided into six sections (a1, a2, b, c, d, and e). a1: larger part of the outermost circle representing miRNA expression; a2: colored rectangles representing mRNA transcripts (as labeled); b: scatterplots illustrating the Ct value (ranging 6.5 to 39; greater values are closer to the center) for the respective microRNA or mRNA. The glyphs are color coded with greater Ct values (low expression) in red, while smaller Ct values (high expression) are in green (median of the data = 22.5; thin white line). c: line plots presenting the z-scores of the transcript expression data (ranging from −3.4 to 3.5; smaller values are closer to the center of circle, the calculation of the z-scores is as described in the methods section), and d: histogram plots showing the log10 transformed relative abundance (to insulin-negative tissues, as described in the methods section in detail), ranging from −7.3 to 10.7. Lesser values are closer to the center of circle. The thin gray lines are placed at every 5% of the data. Section "e" is the innermost plot which displays correlation between miRNAs and the seven mRNAs (Pearson's r > 0.6). The correlations are plotted in the form of colored link (as per the labeled colors of the mRNA).
(E) Bootstrapping involves random deletion (red dots) of multiple samples and replacing them with duplicates from the remainder set (green dots). Each iteration provides a new set, thus eliminating sampling bias.
(F) Frequency/importance (%) of microRNAs (X axis) detected following penalized logistic regression and 1000 machine learning/resampling validation (bootstraps). Highlighted top eight microRNAs represent those common to logistic and linear regression analyses
(G) A t-SNE plot of discovery set using the top eight microRNAs (H) or random eight miRNAs with similar distribution. Each point represents a sample from the discovery set, while the color of the point indicates tissue type.
See also Figure S1 and Tables S1, S2, and S3.

samples were randomly eliminated at each iteration, while samples from the remaining set were duplicated to maintain the same number of samples at each step (Efron and Tibshirani, 1997; Chernick and Labudde, 2011). After 1000 iterations of resampling, a frequency plot representing the number of times a specific microRNA was found to be highly associated with (pro-)insulin expression was derived (Figure 1F). Such a resampling validation (bootstrapping) procedure eliminates sampling bias. Penalized logistic regression models offer variable (microRNA) selection based on categorical assessment (presence or absence) of (pro-)insulin transcript, while penalized linear regression models facilitate the prediction of (pro-)insulin transcript levels (Goeman, 2010). These analyses (Figure S1B) identified the top eight microRNAs from logistic regression analyses to be common with linear regression outputs (Figure S1C). Although all of the microRNAs derived through our analysis based on the machine learning workflows (Figures 1F and S1B) are important for human (pro-)insulin transcription, we focused our analysis on the most important top eight microRNAs (Figure 1F).

Dimensionality reduction analyses (t-distributed stochastic neighbor embedding; Figure 1G) using the top eight microRNAs from the logistic regression analyses could segregate the insulin-positive islet samples from the insulin-negative tissues. However, three other random sets of eight different microRNAs could not efficiently segregate islet samples from other insulin-negative tissues (Figure 1H and Table S3). In summary, these studies identified a set of microRNAs that are associated with (pro-)insulin gene expression in human islets.

Levels of insulin-associated microRNAs correlate with (pro-)insulin in human tissues that naturally transcribe the insulin gene

Apart from pancreatic islets and thymus, the gallbladder (Choi et al., 2019; Sahu et al., 2009; Dutton et al., 2007) and brain (Devaskar et al., 1994; Mehran et al., 2012; Heller et al., 2010) are known to (naturally) transcribe the insulin gene. We observed insulin, glucagon/somatostatin immunopositivity in majority of the human gallbladder and brain samples that we have assessed (Figure 2A). These samples contained detectable levels of (pro-)hormone gene transcripts (*INS*, *GCG*, and *SST*; Figure 2B). We also measured the abundance of pancreatic transcription factors (*MAFA*, *NGN3*, *HES1*, and *PDX1*) in these islet (n = 53), gallbladder (n = 75), brain (n = 14), and insulin-negative tissue samples (n = 55; endothelial cells, muscle, liver, skin, and spleen; Figure 2B). All data are presented as normalized Ct value measured by TaqMan-based real-time qPCR. As expected, the level of (pro-)insulin gene transcripts in the pancreatic islets (mean Ct value ± SD; 15.74 ± 0.74) was significantly higher than that in the gallbladder (mean Ct value ± SD; 36.72 ± 2.24, $p < 0.0001$ vs islet) or in the brain (mean Ct value ± SD; 38.50 ± 0.96, $p < 0.0001$). We measured the absolute copy number of (pro-)insulin gene transcripts using digital droplet PCR (Maynard et al., 2019) (Figure 2C), which confirmed our qPCR findings (Figure 2B) that the number of *INS* mRNA copies was around a hundred thousand-fold higher in human islets than in human gallbladders and up



miRNA	Specificity	Sensitivity	AUC
miR-183-5p	1.000	0.870	0.920
miR-375-3p	1.000	1.000	1.000
miR-216b-5p	1.000	1.000	1.000
miR-183-3p	1.000	1.000	1.000
miR-7-5p	1.000	0.957	0.966
miR-217-5p	0.852	1.000	0.971
miR-7-2-3p	1.000	1.000	1.000
miR-429-3p	0.963	0.957	0.994
Combined (all eight)	1.000	1.000	1.000

Figure 2. Validation and prediction of insulin transcript-associated microRNAs

(A) Immunostaining of insulin (green), glucagon (red), and/or somatostatin (pink) in freshly isolated/cultured human islet, gallbladder epithelium, and brain neurospheres. Nuclei (DNA) are shown in blue. Scale bar is 20 µm.

(B) Real-time TaqMan qPCR expression profile for pancreatic islet (pro-) hormones and transcription factors in human islets (n = 53), gallbladders (n = 75), brains (n = 14), and the insulin-negative tissues (endothelial cells, muscle, liver, skin, and spleen; n = 55). Results are presented as cycle threshold (Ct) values normalized to 18S rRNA. As a lower Ct value represents the higher abundance of gene transcripts, the Y axis is reversed. Each dot within the polygons represents a different sample. The horizontal solid red line represents the median, the horizontal black dotted line within each polygon represents quartiles, and the polygons represent the density of individual data points and extend to min/max values.

(C) Absolute copies for (pro-)insulin gene transcripts in human islet (n = 11), gallbladder (n = 63), brain (n = 14), endothelial (n = 2), muscle (n = 4) and spleen (n = 8) samples, as measured by digital droplet (dd)PCR. Each dot in the floating bars represents a different sample. Horizontal solid lines represent the mean and boundaries extend to min/max values.

(D) MicroRNA abundance of the top eight microRNAs (from the logistic regression analysis, Figure 1F) in human islets (I; n = 30), gallbladders (G; n = 50); brains (B; n = 4), and spleens (S, n = 34) is presented in an aligned dot plot with mean and SEM. Significant differences were computed using one-way analysis of variance (ANOVA) with Dunn's multiple comparisons test. *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$, ****: $p < 0.0001$; compared to islets.

(E) Receiver operator characteristic (ROC) curve analysis for individual and combined top eight microRNAs obtained through the penalized logistic regression and bootstrap analysis. A separate validation set (n = 50) of human islets (n = 23) and insulin-negative samples (n = 27; consisting of human endothelial cells n = 14, gallbladder n = 11, and muscle n = 2) was used. Area under the curve (AUC), sensitivity, and specificity are shown.

(F) Correlation plot between the Ct value for laboratory-measured (pro-)insulin mRNA (X axis) and the computationally predicted (pro-)insulin Ct value (Y axis) from the islets within the validation set (n = 23), impure islet fraction (n = 85), and pancreas (n = 62) from the prediction set (Spearman r = 0.6, $p < 0.0001$). Each dot represents a different sample, and the color of the point indicates the tissue type in the correlation plot.

See also Figure S1 and Tables S4 and S5.

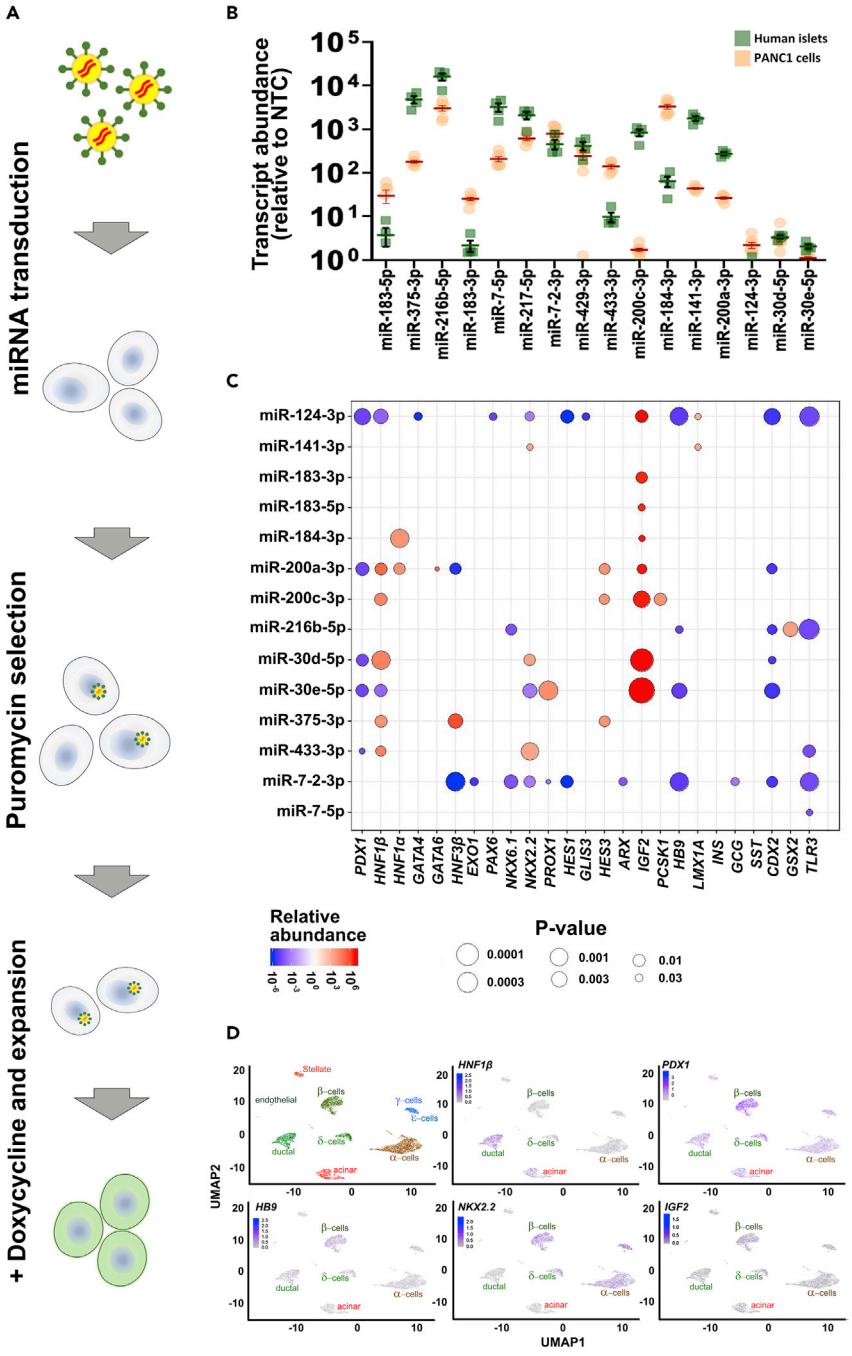


Figure 3. Pancreatic gene expression analysis in microRNA-overexpressing PANC1 lines

(A) A schematic presentation for generating microRNA overexpressing cell lines using lentiviral-mediated, doxycycline-inducible overexpression system.

(B) MicroRNA abundance in human islets (green, n = 4) and microRNA-overexpressing PANC1 cell lines (orange, n = 6 separate experiments) was calculated relative to the respective microRNA levels in non-targeting/scramble control (NTC) PANC1 cell line. A total of 16 microRNAs (presented on X axis) were individually overexpressed in PANC1 cells using doxycycline-inducible lentiviral vectors. The microRNA data are presented as scatter dot plots with mean ± SEM. Each green dot in the plot represents a different islet sample, and each orange dot presents a different preparation/experiment.

Figure 3. *Continued*

(C) Categorical bubble plot displaying effect of overexpression of microRNA (listed on Y axis) as either significant increase (red) or significant decrease (blue) in the expression of a specific transcription factor or (pro-)hormone (presented on X axis) in PANC1 cells. The absence of a circle across any of the pancreatic gene (e.g. *INS* and *SST*) indicates a non-significant effect of corresponding microRNA overexpression. The size of the circles represents the p value ($p < 0.05$, $n = 6$ separate experiments) based on a two-tailed Welch's t-test. Values are presented in relative abundance compared to NTC.

(D) UMAP: Uniform Manifold Approximation and Projection plots generated using Seurat (version 3.2.2) from pancreatic single-cell RNA sequencing (Panc8) data set (SeuratData version 0.2.1). Single-cell data set from four of the five technologies (see methods for details) in Panc8 is presented here. The single cells were clustered into different pancreatic cell types, and expression levels for *HNF1β*, *PDX1*, *HB9*, *NKX2.2*, and *IGF2* across these 13 clusters are highlighted in individual UMAP plots. Only nine of the most relevant clusters are labeled. Data are presented in log-transformed normalized values and are overlaid in shades of blue. Dark blue color shows higher expression in the specific cell clusters, while gray represents absence/low expression of the demonstrated pancreatic gene in these cells.

See also Figure S2 and Table S6.

to a million-fold higher than that in the brain. As anticipated, (pro-)insulin gene transcripts were not detectable in insulin-negative tissues (Figures 2B and 2C).

We then quantified the abundance of the top eight microRNAs identified earlier (green-shaded area in Figure 1F) across human islets, gallbladders, brains, and spleen samples. A significantly lower level of expression was seen for all of the top eight microRNAs ($p < 0.001$; Figure 2D) in gallbladders, brains, and spleens (compared to islets). The levels of all eight microRNAs positively correlated with (pro-)insulin transcript (Table S4) across the three tissue types that naturally produce/transcribe insulin and the insulin-negative (spleen) tissues. These data indicate a potential interdependence of the most important microRNAs and (pro-) insulin gene expression.

Insulin-associated microRNAs are predictive of insulin gene expression

Although insulin-associated microRNAs were able to identify insulin-producing cells in the discovery sample set (Figure 1G), we interrogated their predictive potential in different sets of samples. We generated receiver operating characteristic curves for 50 samples (from the validation set; $n = 23$ islets vs $n = 27$ insulin-negative tissues; Table S1) to test the specificity and sensitivity of these microRNAs in classifying the presence or absence of (pro-) insulin expression. Interestingly, each of these microRNAs demonstrated high specificity (85–100%) and sensitivity (87–100%) in discriminating insulin-positive vs insulin-negative tissues, with an area under the curve of more than 92% (Figure 2E). A combined model, testing all of the top eight miRNAs, returned the highest predictive power (Figure 2E). Our results validate the role of these microRNAs in accurately stratifying insulin-producing tissues in an independent validation set of primary human tissue samples. We then used the prediction set ($n = 202$) of human tissues that naturally express (pro-) insulin gene transcripts at varying levels to further elucidate the predictive power of these microRNAs. We derived the coefficients from the penalized linear regression analysis (Figure S1B) to obtain the odds ratios that constituted a mathematical formula to predict the level (qPCR Ct value) of (pro-)insulin gene expression. Two wet-lab biologists independently carried out the measurements of microRNAs and (pro-) insulin mRNA on the prediction set of human tissue samples consisting of the pancreas, impure islet fractions, gallbladder, and brain ($n = 202$; Table S1). A de-identified list containing the microRNA qPCR Ct values was provided for the computational prediction of (pro-)insulin Ct values. We observed that the levels of these microRNAs could reliably predict the real-time qPCR (laboratory-measured) (pro-)insulin Ct value (Spearman $r = 0.8$, $p < 0.0001$, $n = 202$) in the prediction set of samples (Figure S1D). When the pancreatic tissue was independently analyzed, a significant correlation (Spearman $r = 0.6$, $p < 0.0001$, $n = 170$) between the computationally predicted and laboratory-measured (pro-)insulin Ct values was observed (Figure 2F). In summary, the levels of bootstrapped microRNAs (including top eight microRNAs; Figure S1C) separated categories (Figure 2E) as well as computationally predicted (pro-)insulin Ct values (Figures 2F and S1D). Together, our data demonstrated that the top eight (pro-)insulin-associated microRNAs identified using machine learning workflows were not only associated with (pro-)insulin expression but also predictive of (pro-)insulin transcript abundance in separate sets of primary human tissues.

Forced expression of insulin-associated microRNAs alters pancreatic gene expression

Since associations do not prove causality, we investigated if regulated expression of insulin-associated microRNAs in human pancreatic progenitor cells can promote (pro-) endocrine gene expression. In order to

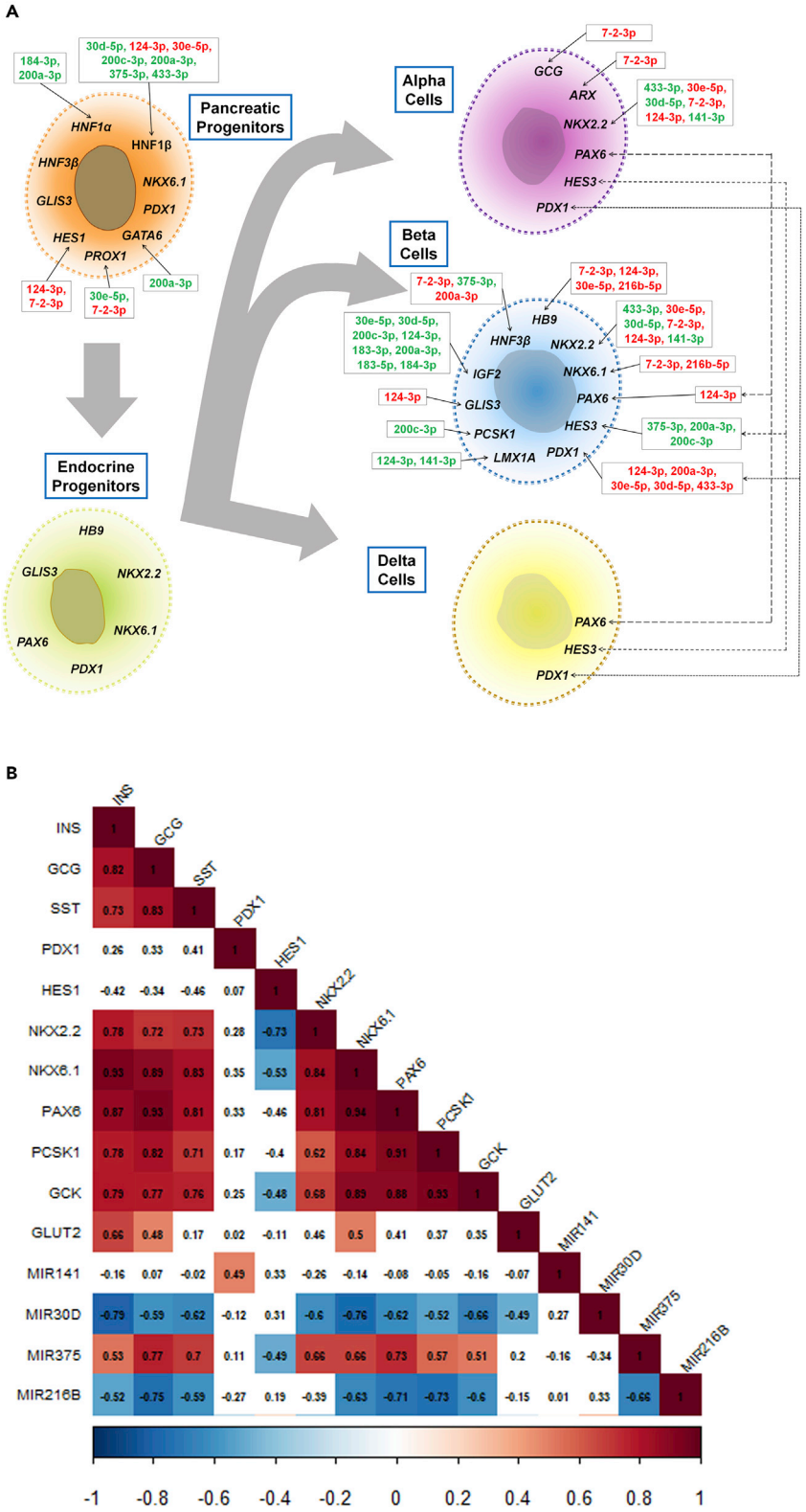


Figure 4. A microRNA regulatory map of human pancreatic cell fates

(A) A map of microRNA regulation was generated based on our data (see Figure 3) of microRNA overexpression in human pancreatic duct cells and the regulation of a select set of lineage-enriched (α-, β-, or δ-) genes known to be important in pancreas development. Key pancreatic genes and transcription factors for a particular developmental stage are shown.

Figure 4. *Continued*

The microRNA labels are colored in green or red based on our data confirming a statistically significant ($p \leq 0.05$) change in gene expression (shown by arrows that each microRNA points to). MicroRNAs that significantly increase gene transcript levels are in green color and those that significantly decrease gene expression are in red. All microRNAs that significantly regulate the expression of a specific gene are listed based on the p value (calculated using a two-tailed Welch's t-test), with the miRNAs at the end of the set being closest to the statistical cutoff value of $p = 0.05$. Since no single microRNA was seen to significantly regulate *INS* or *SST* gene transcript levels, these genes are not listed in the β- and δ-cell lineage.

(B) Correlation between representative microRNAs (miR-141, miR-30d, miR-375, and miR-216b) and pancreatic endocrine transcripts. Correlation analysis was performed using the human islet bulk RNA-seq data set (GSE134068; n = 18). The Pearson correlation r coefficients are presented in black text in each square. Only significant correlations (Pearson $p \leq 0.05$) are presented with a colored fill (depicted in the scale bar) displaying their correlation coefficient.

address this question, we forced the expression of individual microRNAs in human pancreatic progenitor cells (Figure S2A) using a puromycin-selective doxycycline-inducible lentiviral vector system (Figure 3A). Experiments with individual microRNA vectors were carried out as attempts to co-transduce cells with all of the top eight microRNAs resulted in high mortality. Since we were interested in assessing the potential of these microRNAs in differentiation of progenitor or precursor cells (rather than enhancing insulin expression in β-cell lines such as EndoC-βH1), we selected human islet-derived progenitor cells (hIPCs) and commercially available human pancreatic duct (PANC1) cells (Figure S2A) for these studies. A total of 16 different microRNA overexpression and one non-targeting/scramble control vectors were used to generate microRNA-overexpressing lines (Figure S2A). These comprised 14 different microRNAs from our penalized logistic regression bootstrap analysis, including the top eight (Figure 1F). Two other microRNAs (miR-30e-5p and -30d-5p) were also selected for these overexpression studies as they have been previously reported by us (Joglekar et al., 2009c) and others (Tang et al., 2009) in modulating (pro-)insulin gene expression. We first confirmed that each of the 16 PANC1 lines generated for this study demonstrated overexpression of the relevant microRNAs at levels closer to those detected in human islets (Figure 3B). Then, using TaqMan real-time qPCR we assessed the ability of each of these 16 microRNAs to instill a measurable and/or significant change in the expression levels of a select set of 51 human pancreatic mRNA/gene transcripts (Table S6) known to be important in pancreas development/function (Jennings et al., 2015; Chakrabarti and Mirmira, 2003). Only 23 of these 51 mRNAs (shown in Figure 3C) were significantly altered following overexpression of individual microRNAs. Surprisingly, two microRNA-overexpressing lines (miR-217-5p and miR-429-3p) did not significantly change the expression of any of these 51 mRNAs assessed. On the other hand, three or more microRNA-overexpressing lines induced a significant increase in *IGF2*, *HES3*, *NKX2.2*, and *HNF1β* levels compared to non-targeting/scramble control (Figure 3C). At the same time, the expression of *PDX1*, *CDX2*, *TLR3*, and *HB9* was significantly reduced in several other microRNA-overexpressing lines (Figure 3C). Interestingly, certain microRNAs including miR-7-2-3p and miR-7-5p were responsible for significant reduction of pancreatic genes, whereas overexpression of miR-375-3p, miR-183-5p, miR-183-3p, miR-184-3p, miR-141-3p, and miR-200c-3p only led to a significant increase in pancreatic gene expression. None of the microRNA-overexpressing lines induced any significant change in (pro-)insulin and *SST* gene expression. We also mixed PANC1 lines overexpressing each of the top eight (pro-)insulin-associated microRNAs in equal proportion to understand if this would lead to (pro-)insulin expression. A modest (2.2-fold) but insignificant ($p = 0.4$) increase in (pro-)insulin gene transcripts was observed along with significant and more robust increase in the levels of other (pro-)hormone and transcription factors including *GCG*, *NGN3*, *MAFA*, *PAX2*, *PAX6*, *NKX2.2*, and *UCN3* (Figure S2B).

To identify the pancreatic cell type that each of these miRNAs promoted, we analyzed the publicly available pancreatic single-cell sequencing (scSeq) data sets (Butler et al., 2018; Stuart et al., 2019). As expected, these human pancreatic scSeq analyses indicated that *NKX2.2* was predominantly localized to all endocrine pancreatic (α-, β-, δ-, γ-, ε-) cell lineages (Figure 3D). On the other hand, *IGF2*, one of the most common and significantly upregulated genes (Figure 3C), is localized to human pancreatic β cells (Figure 3D). These data suggested that overexpression of microRNAs such as miR-433-3p significantly increased abundance of *HNF1β* and *NKX2.2*, which are localized to endocrine pancreatic cells (Figures 3C and 3D). We then created a microRNA regulatory map for different cell types observed during human pancreas development (Figure 4A). We identified a set of genes that have been previously reported in pancreas development (reviewed in (Jennings et al., 2015; Chakrabarti and Mirmira, 2003)) to define endocrine pancreatic sub-types. We then applied our observations of single microRNA overexpression studies in regulating these pancreatic genes (Figure 3C) to construct an mRNA-microRNA interaction map (Figure 4A). Here, microRNAs are ranked based on the level of significant change (p value) that each of the microRNAs induced

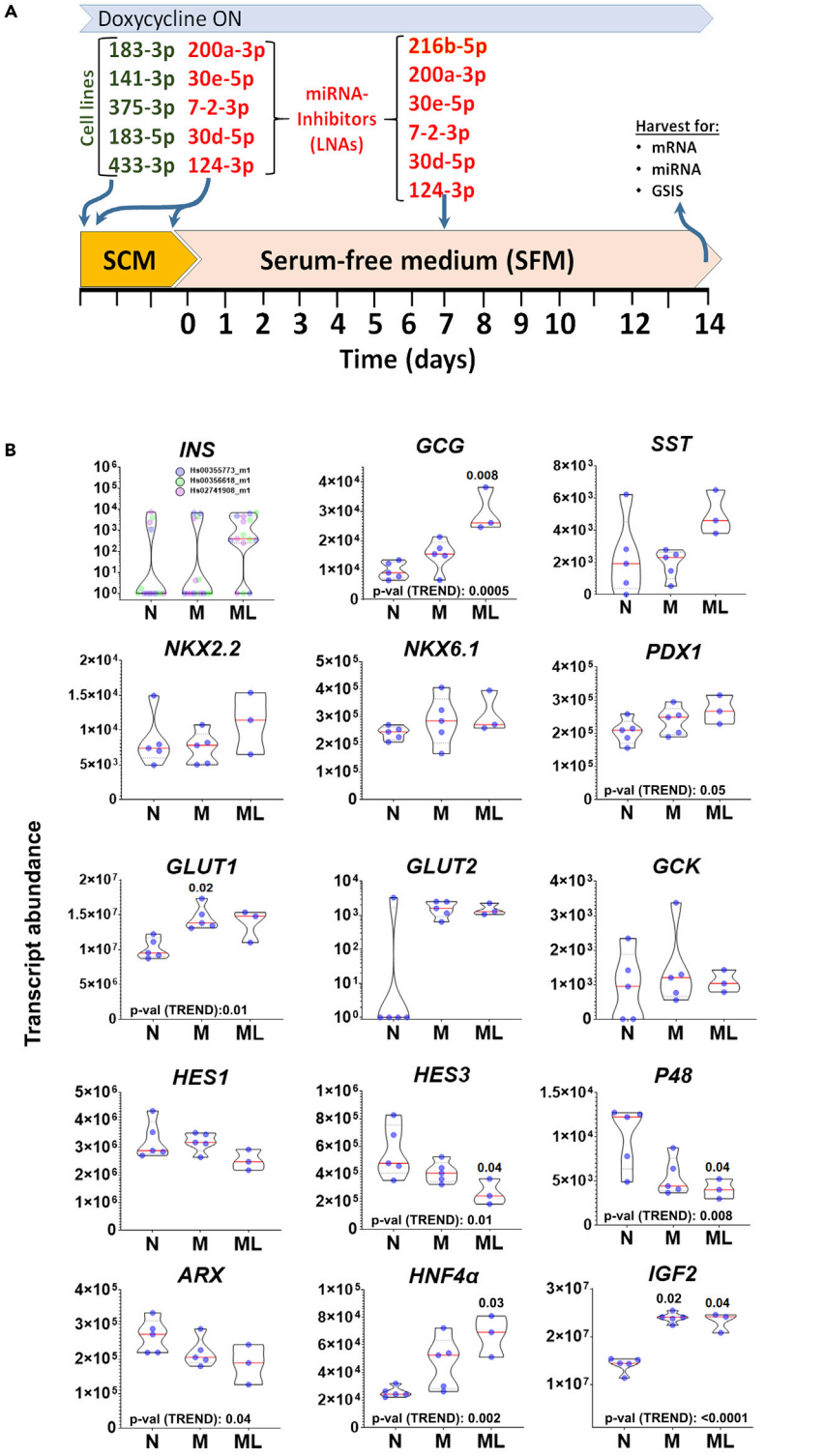


Figure 5. A microRNA overexpression and inhibition model of human pancreatic duct cell differentiation

(A) An outline of the microRNA overexpression and inhibition study to assess the potential combination of microRNAs promoting differentiation to endocrine pancreatic lineage. A non-targeting control microRNA overexpressing line (N) or five selected individual microRNA-overexpressing lines (shown in green) combined in equal proportions in the absence

Figure 5. *Continued*

(M) or presence (ML) of LNA power inhibitors targeting microRNAs (shown in red) were differentiated (see methods). All cells were maintained in serum-containing (growth promoting) media for three days before inducing differentiation in a serum-free medium (SFM), with doxycycline-regulated microRNA expression at all times. Differentiation promoted aggregation and formation of islet-like cell aggregates in serum-free medium. LNA power inhibitors were added again at day seven of the differentiation process in the ML group. Islet-like cell aggregates were harvested for downstream mRNA and microRNA analysis, as well as glucose-stimulated insulin secretion (GSIS) assays at the end of 14 days.

(B) Violin plots displaying transcript abundance of select pancreatic transcription factors, genes, and (pro-) hormones in differentiated human pancreatic duct cells. Transcript abundance was calculated using the fold-over detectable method (Hardikar et al., 2014). On the x axis are data for the non-target/scramble control (N) (n = 5), the combination of five selected individual microRNA-overexpressing cell lines without LNA inhibitors (M) (n = 5), or with LNA inhibitors (ML) (n = 3). Each data point represents a different experiment. The horizontal red line marks the median and dotted black lines within each polygon represent quartiles. All polygons represent the density of individual data points and extend to min/max values. Three different primer probes used to measure the (pro-)insulin gene are shown in three different colors with the TaqMan microRNA assay ID listed in Table S5 and S6. Statistical significance was calculated compared to N using Dunn's multiple comparisons test. Ordinary one-way ANOVA was used for the trend analysis between N, M, and ML. Exact p values for comparison and trend are reported for significantly different comparisons and trend ($p \leq 0.05$). See also Table S6.

to either increase (shown in green) or decrease (shown in red) target mRNA gene expression. As an alternate validation, we assessed the expression levels of these microRNAs and islet mRNAs in our bulk RNA-seq data sets (GSE134068; n = 18 (Wong et al., 2019)) and correlated selected microRNAs (miR-375, miR-141, miR-30d, and miR-216b) with key pancreatic gene transcripts. As shown in (Figure 4B), several pancreatic genes positively correlated with miR-375, while negatively correlating with the levels of miR-216b (or miR-30d) in human islets. These analyses provided additional evidence in the selection of microRNA candidates for future overexpression and knockdown studies. In summary, forced expression of specific microRNAs induced pancreatic genes (Figures 3C, 4A, and 4B) that map to sub-populations of human pancreatic hormone-producing cells (confirmed through scRNA sequencing; Figure 3D) and microRNA-mRNA expression in human islets provided supportive evidence to identify candidates for future studies.

A combination of microRNAs directs differentiation to endocrine pancreatic lineage

Since islets are heterogeneous cell clusters, we hypothesized that protocols involving co-culture, aggregation (and differentiation) of multiple microRNA-overexpressing lines would help in assessing the role of select sets of microRNAs. We previously reported the potential of human pancreatic progenitor cells (PANC1 [Hardikar et al., 2003] and hIPCs [Gershengorn et al., 2004]) to form cell aggregates that contain several transcripts of mature islet-like cells. Our microRNA overexpression data (Figure 3C) identified that certain microRNAs (miR-141-3p, miR-183-3p, miR-433-3p, and miR-183-5p) promote the expression of *IGF2* and *NKX2.2*, which are characteristic of human pancreatic endocrine cells (Figure 3D). Since miR-375-3p has been well known to be an important islet-enriched microRNA and has been observed to promote early pancreatic genes (*HNF1β* and *HNF3β*), we selected it along with the above four microRNA over-expressing lines (miR-141-3p, miR-183-3p, miR-433-3p, and miR-183-5p) in further studies. Other microRNAs (miR-30d-5p, miR-30e-5p, miR-200a-3p, and miR-124-3p) suppressed the expression of the pancreatic master regulator transcription factor *PDX1*. The microRNA miR-7-2-3p was seen to reduce several (pro-)endocrine gene transcripts such as *NKX2.2*, *NKX6.1*, as well as the (pro-)hormone *GCG*. The five microRNAs (miR-30d-5p, miR-30e-5p, miR-200a-3p, miR-124-3p, and miR-7-2-3p) were therefore inhibited in further studies. Since miR-216b-5p was observed to suppress *NKX6.1* that is known to control a gene regulatory network for maintaining human islet β-cell identity (Schaffer et al., 2013), we inhibited miR-216b-5p at later stages of our differentiation protocol (Figure 5A). We used an equal proportions of five different microRNA-overexpressing lines (miR-183-5p, miR-183-3p, miR-375-3p, miR-433-3p, and miR-141-3p), in doxycycline-containing media for 3 days before induction of differentiation in serum-free defined media (Figure 5A). We then inhibited the expression of other microRNAs (miR-7-2-3p, miR-124-3p, miR-200a-3p, miR-30e-5p, miR-30d-5p, miR-216b-5p) using Locked Nucleic Acid (LNA) inhibitors in a stage-specific manner (Figure 5A). We assessed changes in expression of pancreatic genes in non-targeting control/NTC line (referred to as "N"; Figure 5B), a mix of the five selected miRNA-overexpressing cell lines (referred to as "M"; Figure 5B) or the mix of these five miRNA-overexpressing lines with LNA inhibitors (referred to as "ML"; Figure 5B). MiRNA, mRNA, and insulin release/content assessments were carried out for all the three experimental conditions (N, M, and ML).

Combined microRNA overexpression alone (M) or with LNAs (ML) induced a significant increase in *GCG*, *GLUT1*, *HNF4α*, and *IGF2* transcripts (Figure 5B), as compared to the non-targeting control (N). A significant trend was observed for *PDX1*, *HNF4α*, *IGF2*, and *GLUT1*, suggesting the added roles of microRNAs (M) and with LNAs (ML) in regulating their expression. Interestingly, *HES1* and *P48* were reduced in ML, suggesting progression to endocrine (rather than exocrine) lineage. (Pro-)insulin expression was comparable in all the three conditions but was consistently higher when five different microRNA-overexpressing lines and six LNA inhibitors were combined (ML) during differentiation (Figure 5B). Although differentiated control (N) cells showed comparable C-peptide concentrations (20.0 ± 0.7 pmol/L/μg of DNA; mean ± SEM, n = 3) to that in the microRNA overexpressing lines (15.5 + 5.1 pmol/L/μg of DNA; mean ± SEM, n = 3), the latter demonstrated over 4-fold efficient glucose-stimulated insulin secretion (0.8 ± 0.6% of C-peptide content in N vs 4.2 + 1.1% of C-peptide content in M; mean ± SEM, n = 3-4/group). No significant difference in C-peptide content was observed between M and ML groups (data not shown).

We further tested if the effects of these (pro-)insulin transcript-associated microRNAs in human pancreatic duct cells could be replicated in primary human islet cells. We used primary hIPCs that we described earlier (Gershengorn et al., 2004; Joglekar and Hardikar, 2012;Joglekar et al., 2009b, 2016) to create lentiviral-mediated, doxycycline-inducible, microRNA-overexpressing hIPC lines (Figure S2A). Limited availability of freshly isolated human islets, as well as the reduced proliferative capacity after lentiviral transduction, were some major constraints in selecting only three candidate microRNAs as against all of the top eight. Compared to the non-targeting control (NTC), miR-375-3p overexpression led to an increase, although insignificant, in *INS*, *SST*, and *PDX1* transcripts (Figure 6A), while *GCG* transcript levels were reduced in all the three microRNA-overexpressing hIPC lines. Our studies in differentiating human pancreatic (duct and islet derived) microRNA-overexpressing lines demonstrate the potential of these microRNAs in promoting the expression of genes that are known to be restricted to specific endocrine pancreatic sub-sets.

Interdependence of (pro-)insulin expression and microRNAs

To understand the regulatory hierarchy and the interdependence between (pro-)insulin transcripts and insulin-associated microRNAs, we first forced the expression of a furin-cleavable insulin gene (Ren et al., 2007) in five biological replicates (five different pancreas donors) of hIPCs. Furin-cleavable insulin was used to enable the processing of (pro-)insulin in the absence of desired levels of (pro-)hormone convertase in undifferentiated hIPCs. Although overexpression of the furin-cleavable insulin led to a 100- to ~50,000-fold increase in its mRNA as compared to the empty vector transduced cells (Figure 6B), no significant changes were observed in the level of expression of any of the bootstrap signature of microRNAs (including the top eight microRNAs; Figure S3A). These data confirmed that although overexpression of insulin-associated microRNAs (Figures 3, 4, and 5) drives (pro-)endocrine gene transcripts, the forced expression of insulin gene does not change the abundance of microRNAs associated with and predictive of (pro-)insulin gene transcription. We then assessed if the loss of these microRNAs had any impact on human (pro-)insulin gene transcription. We obtained freshly isolated human islets and maintained these in a defined Connaught Medical Research Laboratories (CMRL) medium with or without the specific microRNA LNA power inhibitors for the next 3–6 days. As compared to the control, there was 63% (range: 26–98%) reduction in microRNA expression in just three days of incubation, which further decreased to 72% (range: 29–98%) over the next three days (Figure S3B). This decline in human islet microRNAs significantly reduced the expression of (pro-)insulin gene in miR-375-3p, miR-7-5p, miR-183-3p, miR-183-5p, miR-217-5p, or miR-7-2-3p knockdown islet cells as compared to control at day 6. These data support the necessity of these microRNAs in human (pro-)insulin transcription (Figure 6C).

Since selective depletion of microRNAs reduced (pro-)insulin gene transcription under these *in vitro* conditions, we investigated if lower expression of our bootstrap microRNAs were associated with reduced (pro-)insulin transcription in islets from human subjects. As shown earlier (Figure 2D), low levels of the top eight microRNAs were associated with low levels of (pro-)insulin in human gallbladders and brains. It is well recognized that the progressive loss of pancreatic islet β-cell function is a major cause leading to T2D (Weyer et al., 1999; Kahn, 2003). We therefore screened human islets that were isolated from individuals with T2D. The levels of (pro-)insulin transcripts were significantly lower in all (n = 9) T2D islets (Figure S4A) as compared to islet samples (n = 53) from donors without diabetes. Analysis of all 754 discovery microRNAs in these T2D islets confirmed that several microRNAs were significantly differentially expressed (Figure S4B). Intriguingly, the levels of multiple insulin-associated microRNAs that we identified

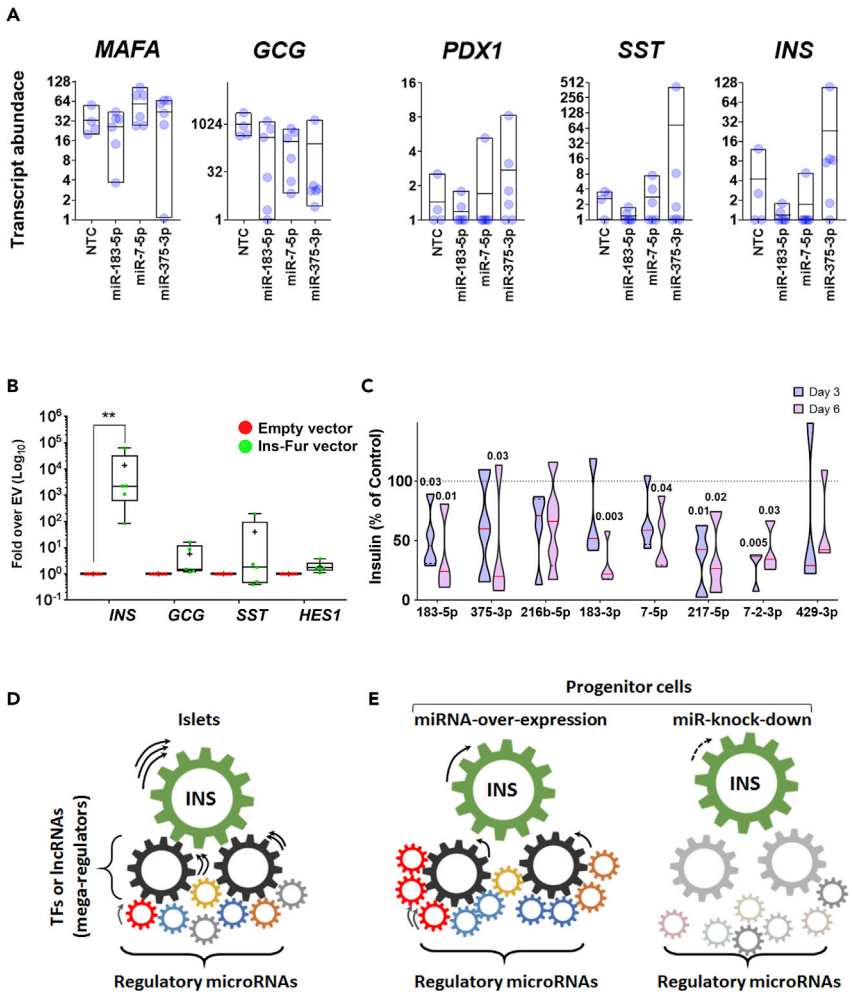


Figure 6. Interdependence of insulin-associated microRNAs and (pro-)insulin gene expression in primary human islet-derived cells

(A) Real-time TaqMan qPCR-based expression analysis of the islet (pro-)hormones and pancreatic transcription factors in microRNA-overexpressing human islet-derived progenitor cells (hIPCs) following *in vitro* differentiation (see methods). Results are normalized to 18S rRNA and presented as transcript abundance, calculated using fold above detectable method (Hardikar et al., 2014). Selected microRNAs (and non-target control/NTC) used in generating overexpressing hIPCs lines are listed on X axis. Each dot in the floating bar plot represents data from 2–3 different islet donors and two separate experiments. The horizontal black line represents the mean and bars extend to min/max values.

(B) Gene expression analysis in hIPCs transduced with the Furin-cleavable (pro-)insulin vector (Ins-Fur; n = 5) or an empty vector (n = 5). Box plots show mean values (indicated by "+"), line at the median and whiskers extending to the minimum and maximum values in the set. Statistical significance based on a two-way ANOVA. **:$p < 0.01$ and five biological (islet donor) preparations of hIPCs.

(C) (Pro-)insulin mRNA expression in human islet cells (from different donors, n = 3–5) following selected power LNA inhibitor-mediated knockdown of microRNAs. The levels of normalized (pro-)insulin gene transcript at day 3 and day 6 are presented as percent of scramble control. The horizontal red line represents the median, and the horizontal dotted black lines within each polygon represent quartiles. The polygons represent the density of distribution of the data and extend to min/max values. Statistical significance was calculated using a two-way ANOVA, and exact p values (compared to control) are presented.

(D and E) A summary of our findings understanding the interdependence of regulatory microRNAs (small colored wheels) and (pro-)insulin (large green wheel) in human islet/pancreatic progenitor cells (see text for details).

See also Figures S3–S5.

in our bootstrap analysis were significantly reduced in T2D islets (Figure S4C). Thus, a pathological reduction in islet (pro-)insulin gene transcript was associated with the depletion of insulin-associated microRNAs, suggesting their potential role in functional β-cell loss in diabetes.

Indeed, these top eight microRNAs regulate several human pancreatic genes (Figure S5A) and target multiple pathways (Figure S5B) related to pancreatic islet development/differentiation (WNT, MAPK, VEGF, TGFβ signaling), insulin production/release (calcium signaling, gap junctions, pancreatic secretion), and islet dysfunction (T1D, T2D, MODY, RNA degradation pathways). Overall, our study results using unbiased machine learning analyses identified a set of microRNAs that are associated with, predictive of, and necessary for (pro-)insulin gene transcription in human islets.

DISCUSSION

Subtle changes in microRNA expression and/or processing can have important biological consequences on (pro-)insulin gene transcription. Conventional genetic manipulation strategies in non-human systems have used knockouts, reporters, and transgenic overexpression models. Such approaches (Lynn et al., 2007; Latreille et al., 2015), although useful to identify and understand the functional role(s) of individual microRNAs, do not fully recapitulate physiologically relevant changes in human pancreatic gene expression networks. The goal of our study was to identify microRNAs associated with human (pro-)insulin gene transcription via profiling human cells and tissues using objective analytical workflows. Machine learning approaches that take out the sampling bias through simulation of a thousand sets of insulin-positive and insulin-negative human tissues were used to derive a signature of insulin-associated microRNAs. We then correlated the abundance of these microRNAs in multiple human tissues representing physiological and pathological depletion of (pro-)insulin, followed by an assessment of their regulatory potential in human pancreatic progenitors and islet cells.

We observed that the overexpression of any single microRNA from our analysis was unable to restore (pro-)insulin expression in pancreatic duct progenitor cells. However, depletion of a single microRNA (such as miR-183-5p, -375-3p, -183-3p, -7-5p, -7-2-3p, or -217-5p) was responsible for a significant reduction but not total absence in (pro-)insulin transcript (Figure 6C). These data suggest that the insulin-associated microRNAs that we identified are not direct regulators of (pro-)insulin gene transcription but possibly act through the coordination of regulatory elements and transcription factors (Figure 6D). Indeed, microRNA target prediction across seven different platforms using miRSystem (Lu et al., 2012) confirmed the absence of any targeting microRNAs for human (pro-)insulin gene transcripts. Our data from human pancreatic progenitor and islet cells demonstrate that overexpression of microRNAs associated with (pro-)insulin abundance (Figure S1C) regulates several pancreatic transcription factors and a combination of these factors fine-tune (pro-)insulin expression, while losses of these miRNAs impact significantly on (pro-)insulin mRNA levels (Figure 6E).

Several microRNAs (including miR-375, miR-7, miR-200, miR-184, miR-124, miR-429) from our bootstrap analysis were previously reported to have specific roles in human β-cell maintenance and function such as (pro-)insulin transcription, glucose uptake, calcium signaling, insulin exocytosis, and β-cell apoptosis (Esguerra et al., 2018; Bolmeson et al., 2011; Wang et al., 2013; Martinez-Sanchez et al., 2018; Filios et al., 2014; Belgardt et al., 2015; Baroukh et al., 2007). In addition to this, we identified several other microRNAs that were not previously associated with human (pro-)insulin transcript levels (e.g. miR-217-5p, miR-183-3p, miR-7-2-3p, miR-183-5p). These microRNAs are, however, reported in other context; miR-217 was noted in developing zebrafish intestine, liver, and pancreas (Stuckenholz et al., 2009) and was shown to inhibit proliferation and invasion of pancreatic adenocarcinoma cells (Chen et al., 2017). MiR-7-2-3p was described in human blood cells following saturated fatty acid feeding (Lopez et al., 2018), while miR-183-5p was reported through *in silico* studies as a T2D gene set targeting microRNA (Baran-Gale et al., 2013).

We also recognize that few previously reported microRNAs were not selected through our analyses. MiR-204, a β-cell-enriched microRNA, was shown to target *MAFA* and block insulin production (Xu et al., 2013, 2016). However, another study demonstrated that although miR-204 is highly abundant in human islets, it did not significantly change (pro-)insulin gene transcript levels in human islets as well as in EndoC-betaH1 (Marzinotto et al., 2017). Another such candidate microRNA is miR-30d, a member of the miR-30 family that we reported to be islet enriched in human fetal islets (Joglekar et al., 2009c). MiR-30d was also reported to regulate (pro-)insulin gene transcription in mouse cells (Tang et al., 2009). In our analyses, both miR-204 and miR-30d were identified as highly significant ($p < 2.349 \times 10^{-19}$) microRNAs and expressed at higher levels in adult human islets (8.4-cycles or $2^{8.4} = 338$-fold higher for miR-204 and 2.5-cycles or $2^{2.5} = 5.7$-fold higher for miR-30d; Table S2). These microRNAs, although islet enriched, were not selected through the unbiased

machine learning analytical workflow that we implemented in the identification of microRNAs associated with presence/abundance of (pro-)insulin transcript.

Forced microRNA overexpression was previously reported to enhance differentiation of human insulin-producing cells, most commonly using miR-375 (Nathan et al., 2015; Lahmy et al., 2014, 2016; Piran et al., 2017; Shaer et al., 2014; Williams et al., 2020). Similarly, miR-7 overexpression increased human embryonic stem (ES) cell differentiation (Lopez-Beas et al., 2018), whereas a cluster of microRNAs (miR-375, let-7a, let-7g, and miR-200a) was overexpressed using a polycistronic construct (Jin et al., 2019) during the differentiation of human ES cells into insulin-producing cells. The microRNAs that were previously reported to induce/improve differentiation to insulin-producing cells (miR-375, miR-200a, miR-7, and miR-9) are observed in our bootstrap analysis (Figure 1F). Interestingly, while our data confirm that miR-375 is significantly abundant in human islets ($p = 1.7 \times 10^{-21}$), miR-9 was identified to be significantly abundant ($p = 0.02$) in human insulin-negative tissues (Table S2). This is in line with the approach of miR-375 overexpression along with miR-9 inhibition, which was reported in the differentiation of human bone-marrow-derived stromal cells to insulin-producing lineage (Jafarian et al., 2015). MicroRNA-7 is reported to target *Pax6* in mouse islet cells and reduce (pro-)insulin expression (Kredo-Russo et al., 2012), corroborating with our observations that miR-7 overexpression (in PANC1 or hIPCs) did not induce (pro-)insulin expression. The modest increase in (pro-)insulin transcripts or C-peptide production observed in our study may be due to the choice of microRNA overexpressing line(s), as well as the timing and dosage of microRNA overexpression or inhibition achieved in the study.

Targeted transient depletion of candidate microRNAs (Figure 6C) in human islets reduced (pro-)insulin expression, validating the previously reported role of microRNAs (Melkman-Zehavi et al., 2011) and miRNA processing (Kalis et al., 2011; Lynn et al., 2007) in reducing (pro-)insulin transcription in rodent models. Indeed, several of these microRNAs (Figure S1C) were significantly reduced in islets from organ donors with T2D, in agreement with data previously published using a smaller set of human islet samples and other (small RNA sequencing) technologies (Kameswaran et al., 2014). Even though knockdown of miR-216b-5p and miR-429-3p did not result in significant reduction of (pro-)insulin transcript, this could be the result of inefficient knockdown or higher interdependence/correlation between some families of microRNAs affecting function. Future studies with multiple replicates of human islet cells to confirm our observations are merited.

Our study strengths reside in the generation of this resource of microRNA expression profiles in a large number and mixture of human tissues using a previously verified sensitive and reliable OpenArray microRNA profiling platform. Compared to other studies on profiling microRNAs in human pancreas/islets (Correa-Medina et al., 2009; Joglekar et al., 2009a; Bolmeson et al., 2011; van de Bunt et al., 2013; Sebastiani et al., 2015; Klein et al., 2013; Grieco et al., 2017; Fred et al., 2010; Kameswaran et al., 2014), our microRNA data present one of the largest human tissue set ($n = 353$) of insulin-positive and insulin-negative tissues. Our design incorporates three separate sets of samples for use in discovery, validation, and prediction studies, as well as machine learning resampling approaches to take out sampling bias. Our study revealed correlations between this specific set of microRNAs and human (pro-)insulin transcripts in other tissues that naturally transcribe the (pro-)insulin gene. To our knowledge, this is the first report demonstrating an association of microRNA levels to (pro-)insulin transcript across a range of insulin-positive and insulin-negative primary human tissues. Our study also demonstrates the potential regulatory role of these microRNAs in human diabetes. Since functional loss of β cells is common to T1D (Atkinson et al., 2014; Kim et al., 2016) as well as T2D, our observations open future avenues for assessing the roles of specific microRNAs in diabetes progression. Another strength is the generation of this resource of microRNA overexpressing human pancreatic progenitor lines that allows the study of microRNA-target interactions. The use of analytical and validation tools led to the identification of microRNAs including miR-217-5p, miR-7-2-3p, miR-183-3p, and miR-183-5p that were not previously associated with human (pro-)insulin gene transcription. Without such an approach, the relevance and importance of these microRNAs in islet cells would not have been deciphered.

Limitations of the study

A limitation is the under-representation of some tissue types within our diverse sample set (Table S1), which was difficult to control due to its human origin. We had to work with available tissue sets while trying to include a diverse array of insulin-positive and insulin-negative tissues available through our collaborating

investigators. The use of machine learning algorithms with resampling validation (Efron and Tibshirani, 1997; Chernick and Labudde, 2011) was therefore useful to eliminate any sampling bias. Since 10,000 re-sampling iterations delivered the same signature as with 1000 iterations, we used the latter to save on computing time. However, we acknowledge that the tissue specificity/enrichment of these microRNAs needs to be further examined using more numbers and different types (e.g. thymus) of insulin-producing tissues. The other study limitation is the focus on a selected set and combination of microRNAs presented in this study. We selected the top eight microRNAs from our logistic regression bootstrap analyses as they were also common to the linear regression workflow (Figure S1C). Since the top eight microRNAs (n = 8) could generate up to 255 different combinations $\left(\sum_{r=1}^{8} \frac{n!}{r!*(n-r)!}\right)$ of microRNA overexpressing cell lines, we were unable to plan for all of these combinations in the current study. Therefore, a single combination of most relevant microRNA lines was planned and analyzed. Our study presents an example (from 255 combinations) and offers this resource of human pancreatic microRNA-overexpressing lines and data sets for future studies. The association of microRNA levels to (pro-)insulin transcripts in human gallbladder and brain is intriguing. Although lower microRNA levels in these tissues do not establish causality, we also observe a significant reduction in several of our bootstrap microRNAs in islets from human T2D donors (with reduced (pro-)insulin levels). Mixing top eight microRNA-overexpressing lines in equal proportion did not significantly increase (pro-)insulin abundance, which suggests that although the machine learning approach identified the important microRNAs, the understanding of the timing and dosage of their expression/repression and the proportion of their individual contribution in regulating insulin transcription is necessary.

METHODS

All methods can be found in the accompanying transparent methods supplemental file.

Resource availability

Lead contact

Anandwardhan A. Hardikar; email: A.Hardikar@westernsydney.edu.au.

School of Medicine, Western Sydney University, 30.2.27, Narellan Road & Gilchrist Drive.

Campbelltown, NSW 2560, Australia.

Materials availability

MicroRNA overexpressing pancreatic duct cell lines can be obtained with reasonable request to the lead contact.

Data and code availability

RNA-seq data used and analyzed in this study were deposited on the Gene Expression Omnibus (GEO) database and are available through the study accession number GSE134068. Our script used for penalized regression and bootstrap analyses carried herein are available through https://github.com/Isletbiology/Penalized-regression.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.102379.

ACKNOWLEDGMENTS

The research presented herein has been funded through grants from the Australian Research Council Future Fellowship (FT110100254), the Juvenile Diabetes Research Foundation (JDRF) Australia T1D Clinical Research Network (JDRF/4-CDA2016-228-MB), and the University of Sydney CDIP grants to A.A.H. A.A.H, L.T.D., and A.E.S. are also supported through visiting professorships from the Danish Diabetes Academy, funded by the Novo Nordisk Foundation, grant number NNF17SA0031406 (2016-18 and 2019-22). W.K.M.W. acknowledges previous support from the Australian Postgraduate Award, University of Sydney, JDRF Australia PhD top-up award, and current funding through JDRF Australia/Helmsley Charitable Trust. M.V.J. was supported through a JDRF USA advanced post-doctoral award (3-APF-2016-178-

A-N) and currently a transition award from JDRF, USA. J.R.G. holds the Wenkart Chair of the Endothelium. P.O.C. was an NHMRC senior practitioner fellow and funded by an NHMRC program grant and the JDRF. The St Vincent's Institute receives support from the Operational Infrastructure Support Scheme of the Government of Victoria. R.C.W.M. acknowledges support from the RGC Theme-based Research Scheme (T12-402/13N) and Research Impact Fund (R4012-18), the Focused Innovation Scheme, and Faculty Post-doctoral Scheme of the Chinese University of Hong Kong. A.E.S. was supported through a Danish Diabetes Academy post-doctoral grant, supported by the Novo Nordisk Foundation. We thank Ms. Cody Lee-Maynard, Ms. Dana AlRijjal, and Dr. Najeeb Syed for assistance in laboratory analytical work. A.A.H. acknowledges interaction(s) with Dr. Khalid Fakhro, Ms. Shihana Fathima, Prof. Alicia J. Jenkins, Prof. Anthony C. Keech, Prof. Val Gebski and infrastructure support from the NHMRC CTC, Faculty of Medicine & Health, University of Sydney; Australia, School of Medicine, Western Sydney University, Australia; Western Sydney University, Ingham Institute, Liverpool; Australia and the Rebecca L. Cooper Medical Research Foundation. The support of all surgical team members contributing to the consenting and acquisition of research tissue samples, all organ donors, and supporting family members is gratefully acknowledged.

AUTHOR CONTRIBUTIONS

Conceptualization, A.A.H.; methodology, A.A.H., W.K.M.W., and M.V.J.; software, W.K.M.W., G.J., G.J.M., C.X.D., A.C., A.S.A., A.S.J., A.E.S., A.S.A., L.T.D., and R.C.W.M.; validation, W.K.M.W., M.V.J., V.S., and A.A.H.; formal analysis, W.K.M.W., M.V.J., V.S., G.J., A.E.S., L.T.D., R.C.W.M., and A.A.H.; investigation, W.K.M.W., M.V.J., V.S., D.G., R.J.F., S.N.S., S.S., T.S., and D.L.; resources, W.K.M.W., M.V.J., V.S., D.G., S.N.S., Y.V.C., B.H., C.K.L., J.H., J.R.G., T.L., T.W.K., H.E.T., P.J.O., G.J.G., D.M., A.M.S., W.J.H., and A.A.H.; data curation, G.J., Y.V.C., W.J.H., and R.C.W.M.; writing – original draft, W.K.M.W., M.V.J., and A.A.H.; writing – review & editing, all authors; visualization, A.A.H.; supervision, A.A.H.; project administration, A.A.H.; funding acquisition, A.A.H.

DECLARATION OF INTERESTS

A patent application (WO2019000017A1) was filed.

Received: September 16, 2020
Revised: February 19, 2021
Accepted: March 29, 2021
Published: April 23, 2021

REFERENCES

Atkinson, M.A., Eisenbarth, G.S., and Michels, A.W. (2014). Type 1 diabetes. Lancet *383*, 69–82.

Baran-Gale, J., Fannin, E.E., Kurtz, C.L., and Sethupathy, P. (2013). Beta cell 5'-shifted isomiRs are candidate regulatory hubs in type 2 diabetes. PLoS One *8*, e73240.

Baroukh, N., Ravier, M.A., Loder, M.K., Hill, E.V., Bounacer, A., Scharfmann, R., Rutter, G.A., and van Obberghen, E. (2007). MicroRNA-124a regulates Foxa2 expression and intracellular signaling in pancreatic beta-cell lines. J. Biol. Chem. *282*, 19575–19588.

Belgardt, B.F., Ahmed, K., Spranger, M., Latreille, M., Denzler, R., Kondratiuk, N., von Meyenn, F., Villena, F.N., Herrmanns, K., Bosco, D., et al. (2015). The microRNA-200 family regulates pancreatic beta cell survival in type 2 diabetes. Nat. Med. *21*, 619–627.

Bolmeson, C., Esguerra, J.L., Salehi, A., Speidel, D., Eliasson, L., and Cilio, C.M. (2011). Differences in islet-enriched miRNAs in healthy and glucose intolerant human subjects. Biochem. Biophys. Res. Commun. *404*, 16–22.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol. *36*, 411–420.

Chakrabarti, S.K., and Mirmira, R.G. (2003). Transcription factors direct the development and function of pancreatic beta cells. Trends Endocrinol. Metab. *14*, 78–84.

Chen, Q., Wang, P., Fu, Y., Liu, X., Xu, W., Wei, J., Gao, W., Jiang, K., Wu, J., and Miao, Y. (2017). MicroRNA-217 inhibits cell proliferation, invasion and migration by targeting Tpd52l2 in human pancreatic adenocarcinoma. Oncol. Rep. *38*, 3567–3573.

Chernick, M.R., and Labudde, R. (2011). An Introduction to Bootstrap Methods with Applications (John Wiley & Sons, Inc).

Choi, S.I., Lee, B., Woo, J.H., Jeong, J.B., Jun, I., and Kim, E.K. (2019). APP processing and metabolism in corneal fibroblasts and epithelium as a potential biomarker for Alzheimer's disease. Exp. Eye Res. *182*, 167–174.

Correa-Medina, M., Bravo-Egana, V., Rosero, S., Ricordi, C., Edlund, H., Diez, J., and Pastori, R.L. (2009). MicroRNA miR-7 is preferentially expressed in endocrine cells of the developing and adult human pancreas. Gene Expr. Patterns *9*, 193–199.

Devaskar, S.U., Giddings, S.J., Rajakumar, P.A., Carnaghi, L.R., Menon, R.K., and Zahm, D.S. (1994). Insulin gene expression and insulin synthesis in mammalian neuronal cells. J. Biol. Chem. *269*, 8445–8454.

Dutton, J.R., Chillingworth, N.L., Eberhard, D., Brannon, C.R., Hornsey, M.A., Tosh, D., and Slack, J.M. (2007). Beta cells occur naturally in extrahepatic bile ducts of mice. J. Cell Sci. *120*, 239–245.

Efron, B., and Tibshirani, R. (1997). Improvements on cross-validation: the .632+ bootstrap method. J. Am. Stat. Assoc. *92*, 548–560.

Esguerra, J.L.S., Nagao, M., Ofori, J.K., Wendt, A., and Eliasson, L. (2018). MicroRNAs in islet hormone secretion. Diabetes Obes. Metab. *20*, 11–19.

Farr, R.J., Januszewski, A.S., Joglekar, M.V., Liang, H., Mcaulley, A.K., Hewitt, A.W., Thomas, H.E., Loudovaris, T., Kay, T.W., Jenkins, A., and Hardikar, A.A. (2015). A comparative analysis of high-throughput platforms for validation of a circulating microRNA signature in diabetic retinopathy. Sci. Rep. *5*, 10375.

Filios, S.R., Xu, G., Chen, J., Hong, K., Jing, G., and Shalev, A. (2014). MicroRNA-200 is induced by thioredoxin-interacting protein and regulates Zeb1 protein signaling and beta cell apoptosis. J. Biol. Chem. *289*, 36275–36283.

Fred, R.G., Bang-Berthelsen, C.H., Mandrup-Poulsen, T., Grunnet, L.G., and Welsh, N. (2010). High glucose suppresses human islet insulin biosynthesis by inducing miR-133a leading to decreased polypyrimidine tract binding protein-expression. PLoS One *5*, e10843.

Gershengorn, M.C., Hardikar, A.A., Wei, C., Geras-Raaka, E., Marcus-Samuels, B., and Raaka, B.M. (2004). Epithelial-to-mesenchymal transition generates proliferative human islet precursor cells. Science *306*, 2261–2264.

Goeman, J.J. (2010). L1 penalized estimation in the Cox proportional hazards model. Biom J. *52*, 70–84.

Grieco, F.A., Sebastiani, G., Juan-Mateu, J., Villate, O., Marroqui, L., Ladriere, L., Tugay, K., Regazzi, R., Bugliani, M., Marchetti, P., et al. (2017). MicroRNAs miR-23a-3p, miR-23b-3p, and miR-149-5p regulate the expression of proapoptotic BH3-only proteins DP5 and PUMA in human pancreatic beta-cells. Diabetes *66*, 100–112.

Hardikar, A.A., Farr, R.J., and Joglekar, M.V. (2014). Circulating microRNAs: understanding the limits for quantitative measurement by real-time PCR. J. Am. Heart Assoc. *3*, e000792.

Hardikar, A.A., Marcus-Samuels, B., Geras-Raaka, E., Raaka, B.M., and Gershengorn, M.C. (2003). Human pancreatic precursor cells secrete FGF2 to stimulate clustering into hormone-expressing islet-like cell aggregates. Proc. Natl. Acad. Sci. U S A. *100*, 7117–7122.

Heller, R.S., Tsugu, H., Nabeshima, K., and Madsen, O.D. (2010). Intracranial ectopic pancreatic tissue. Islets *2*, 65–71.

Jafarian, A., Taghikani, M., Abroun, S., Allahverdi, A., Lamei, M., Lakpour, N., and Soleimani, M. (2015). The generation of insulin producing cells from human mesenchymal stem cells by MiR-375 and anti-MiR-9. PLoS One *10*, e0128650.

Jennings, R.E., Berry, A.A., Strutt, J.P., Gerrard, D.T., and Hanley, N.A. (2015). Human pancreas development. Development *142*, 3126–3137.

Jin, W., Mulas, F., Gaertner, B., Sui, Y., Wang, J., Matta, I., Zeng, C., Vinckier, N., Wang, A., Nguyen-Ngoc, K.V., et al. (2019). A network of microRNAs acts to promote cell cycle exit and differentiation of human pancreatic endocrine cells. iScience *21*, 681–694.

Joglekar, M.V., and Hardikar, A.A. (2012). Isolation, expansion, and characterization of human islet-derived progenitor cells. Methods Mol. Biol. *879*, 351–366.

Joglekar, M.V., Joglekar, V.M., and Hardikar, A.A. (2009a). Expression of islet-specific microRNAs during human pancreatic development. Gene Expr. Patterns *9*, 109–113.

Joglekar, M.V., Joglekar, V.M., Joglekar, S.V., and Hardikar, A.A. (2009b). Human fetal pancreatic insulin-producing cells proliferate in vitro. J. Endocrinol. *201*, 27–36.

Joglekar, M.V., Patil, D., Joglekar, V.M., Rao, G.V., Reddy, D.N., Mitnala, S., Shouche, Y., and Hardikar, A.A. (2009c). The miR-30 family microRNAs confer epithelial phenotype to human pancreatic cells. Islets *1*, 137–147.

Joglekar, M.V., Sahu, S., Wong, W.K.M., Satoor, S.N., Dong, C.X., Farr, R.J., Williams, M.D., Pandya, P., Jhala, G., Yang, S.N.Y., et al. (2021). A pro-endocrine pancreatic transcriptional program established during development is retained in human gallbladder epithelial cells. bioRxiv. https://doi.org/10.1101/2021.03.02.433636.

Joglekar, M.V., Trivedi, P.M., Kay, T.W., Hawthorne, W.J., O'connell, P.J., Jenkins, A.J., Hardikar, A.A., and Thomas, H.E. (2016). Human islet cells are killed by BID-independent mechanisms in response to FAS ligand. Apoptosis *21*, 379–389.

Kahn, S.E. (2003). The relative contributions of insulin resistance and beta-cell dysfunction to the pathophysiology of Type 2 diabetes. Diabetologia *46*, 3–19.

Kalis, M., Bolmeson, C., Esguerra, J.L., Gupta, S., Edlund, A., Tormo-Badia, N., Speidel, D., Holmberg, D., Mayans, S., Khoo, N.K., et al. (2011). Beta-cell specific deletion of Dicer1 leads to defective insulin secretion and diabetes mellitus. PLoS One *6*, e29166.

Kameswaran, V., Bramswig, N.C., Mckenna, L.B., Penn, M., Schug, J., Hand, N.J., Chen, Y., Choi, I., Vourekas, A., Won, K.J., et al. (2014). Epigenetic regulation of the DLK1-MEG3 microRNA cluster in human type 2 diabetic islets. Cell Metab. *19*, 135–145.

Kim, K.W., HO, A., Alshabee-Akil, A., Hardikar, A.A., Kay, T.W., Rawlinson, W.D., and Craig, M.E. (2016). Coxsackievirus B5 infection induces dysregulation of microRNAs predicted to target known type 1 diabetes risk genes in human pancreatic islets. Diabetes *65*, 996–1003.

Klein, D., Misawa, R., Bravo-Egana, V., Vargas, N., Rosero, S., Piroso, J., Ichii, H., Umland, O., Zhijie, J., Tsinoremas, N., et al. (2013). MicroRNA expression in alpha and beta cells of human pancreatic islets. PLoS One *8*, e55064.

Kloosterman, W.P., Lagendijk, A.K., Ketting, R.F., Moulton, J.D., and Plasterk, R.H. (2007). Targeted inhibition of miRNA maturation with morpholinos reveals a role for miR-375 in pancreatic islet development. PLoS Biol. *5*, e203.

Kredo-Russo, S., Mandelbaum, A.D., Ness, A., Alon, I., Lennox, K.A., Behlke, M.A., and Hornstein, E. (2012). Pancreas-enriched miRNA refines endocrine cell differentiation. Development *139*, 3021–3031.

Lahmy, R., Soleimani, M., Sanati, M.H., Behmanesh, M., Kouhkan, F., and Mobarra, N. (2014). MiRNA-375 promotes beta pancreatic differentiation in human induced pluripotent stem (hiPS) cells. Mol. Biol. Rep. *41*, 2055–2066.

Lahmy, R., Soleimani, M., Sanati, M.H., Behmanesh, M., Kouhkan, F., and Mobarra, N. (2016). Pancreatic islet differentiation of human embryonic stem cells by microRNA overexpression. J. Tissue Eng. Regen. Med. *10*, 527–534.

Latreille, M., Herrmanns, K., Renwick, N., Tuschl, T., Malecki, M.T., Mccarthy, M.I., Owen, K.R., Rulicke, T., and Stoffel, M. (2015). miR-375 gene dosage in pancreatic beta-cells: implications for regulation of beta-cell mass and biomarker development. J. Mol. Med. (Berl) *93*, 1159–1169.

Lopez-Beas, J., Capilla-Gonzalez, V., Aguilera, Y., Mellado, N., Lachaud, C.C., Martin, F., Smani, T., Soria, B., and Hmadcha, A. (2018). miR-7 modulates hESC differentiation into insulin-producing beta-like cells and contributes to cell maturation. Mol. Ther. Nucleic Acids *12*, 463–477.

Lopez, S., Bermudez, B., Montserrat-de la Paz, S., Abia, R., and Muriana, F.J.G. (2018). A microRNA expression signature of the postprandial state in response to a high-saturated-fat challenge. J. Nutr. Biochem. *57*, 45–55.

Lu, T.P., Lee, C.Y., Tsai, M.H., Chiu, Y.C., Hsiao, C.K., Lai, L.C., and Chuang, E.Y. (2012). miRSystem: an integrated system for characterizing enriched functions and pathways of microRNA targets. PLoS One *7*, e42390.

Lynn, F.C., Skewes-Cox, P., Kosaka, Y., Mcmanus, M.T., Harfe, B.D., and German, M.S. (2007). MicroRNA expression is required for pancreatic islet cell genesis in the mouse. Diabetes *56*, 2938–2945.

Martinez-Sanchez, A., Nguyen-Tu, M.S., Cebola, I., Yavari, A., Marchetti, P., Piemonti, L., de Koning, E., Shapiro, A.M.J., Johnson, P., Sakamoto, K., et al. (2018). MiR-184 expression is regulated by AMPK in pancreatic islets. FASEB J. *32*, 2587–2600.

Marzinotto, I., Pellegrini, S., Brigatti, C., Nano, R., Melzi, R., Mercalli, A., Liberati, D., Sordi, V., Ferrari, M., Falconi, M., et al. (2017). miR-204 is associated with an endocrine phenotype in human pancreatic islets but does not regulate the insulin mRNA through MAFA. Sci. Rep. *7*, 14051.

Maynard, C.L., Wong, W.K.M., Hardikar, A.A., and Joglekar, M.V. (2019). Droplet digital PCR for measuring absolute copies of gene transcripts in human islet-derived progenitor cells. Methods Mol. Biol. *2029*, 37–48.

Mehran, A.E., Templeman, N.M., Brigidi, G.S., Lim, G.E., Chu, K.Y., Hu, X., Botezelli, J.D., Asadi, A., Hoffman, B.G., Kieffer, T.J., et al. (2012). Hyperinsulinemia drives diet-induced obesity independently of brain insulin production. Cell Metab. *16*, 723–737.

Melkman-Zehavi, T., Oren, R., Kredo-Russo, S., Shapira, T., Mandelbaum, A.D., Rivkin, N., Nir, T., Lennox, K.A., Behlke, M.A., Dor, Y., and Hornstein, E. (2011). miRNAs control insulin content in pancreatic beta-cells via downregulation of transcriptional repressors. EMBO J. *30*, 835–845.

Nathan, G., Kredo-Russo, S., Geiger, T., Lenz, A., Kaspi, H., Hornstein, E., and Efrat, S. (2015). MiR-

375 promotes redifferentiation of adult human beta cells expanded in vitro. PLoS One *10*, e0122108.

Piran, M., Enderami, S.E., Piran, M., Sedeh, H.S., Seyedjafari, E., and Ardeshirylajimi, A. (2017). Insulin producing cells generation by overexpression of miR-375 in adipose-derived mesenchymal stem cells from diabetic patients. Biologicals *46*, 23–28.

Poy, M.N., Eliasson, L., Krutzfeldt, J., Kuwajima, S., MA, X., Macdonald, P.E., Pfeffer, S., Tuschl, T., Rajewsky, N., Rorsman, P., and Stoffel, M. (2004). A pancreatic islet-specific microRNA regulates insulin secretion. Nature *432*, 226–230.

Poy, M.N., Hausser, J., Trajkovski, M., Braun, M., Collins, S., Rorsman, P., Zavolan, M., and Stoffel, M. (2009). miR-375 maintains normal pancreatic alpha- and beta-cell mass. Proc. Natl. Acad. Sci. U S A *106*, 5813–5818.

Ren, B., O'brien, B.A., Swan, M.A., Koina, M.E., Nassif, N., Wei, M.Q., and Simpson, A.M. (2007). Long-term correction of diabetes in rats after lentiviral hepatic insulin gene therapy. Diabetologia *50*, 1910–1920.

Sahu, S., Joglekar, M.V., Dumbre, R., Phadnis, S.M., Tosh, D., and Hardikar, A.A. (2009). Islet-like cell clusters occur naturally in human gall bladder and are retained in diabetic conditions. J. Cell Mol. Med. *13*, 999–1000.

Schaffer, A.E., Taylor, B.L., Benthuysen, J.R., Liu, J., Thorel, F., Yuan, W., Jiao, Y., Kaestner, K.H., Herrera, P.L., Magnuson, M.A., et al. (2013). Nkx6.1 controls a gene regulatory network required for establishing and maintaining pancreatic Beta cell identity. PLoS Genet. *9*, e1003274.

Sebastiani, G., Po, A., Miele, E., Ventriglia, G., Ceccarelli, E., Bugliani, M., Marselli, L., Marchetti, P., Gulino, A., Ferretti, E., and Dotta, F. (2015). MicroRNA-124a is hyperexpressed in type 2 diabetic human pancreatic islets and negatively regulates insulin secretion. Acta Diabetol. *52*, 523–530.

Shaer, A., Azarpira, N., Vahdati, A., Karimi, M.H., and Shariati, M. (2014). miR-375 induces human decidua basalis-derived stromal cells to become insulin-producing cells. Cell Mol. Biol. Lett. *19*, 483–499.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M.,3R.D., Hao, Y., stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive integration of single-cell data. Cell *177*, 1888–1902 e21.

Stuckenholz, C., Lu, L., Thakur, P., Kaminski, N., and Bahary, N. (2009). FACS-assisted microarray profiling implicates novel genes and pathways in zebrafish gastrointestinal tract development. Gastroenterology *137*, 1321–1332.

Tang, X., Muniappan, L., Tang, G., and Ozcan, S. (2009). Identification of glucose-regulated miRNAs from pancreatic {beta} cells reveals a role for miR-30d in insulin transcription. RNA *15*, 287–293.

van de Bunt, M., Gaulton, K.J., Parts, L., Moran, I., Johnson, P.R., Lindgren, C.M., Ferrer, J., Gloyn, A.L., and Mccarthy, M.I. (2013). The miRNA profile of human pancreatic islets and beta-cells and relationship to type 2 diabetes pathogenesis. PLoS One *8*, e55272.

Wang, Y., Liu, J., Liu, C., Naji, A., and Stoffers, D.A. (2013). MicroRNA-7 regulates the mTOR pathway and proliferation in adult pancreatic beta-cells. Diabetes *62*, 887–895.

Weyer, C., Bogardus, C., Mott, D.M., and Pratley, R.E. (1999). The natural history of insulin secretory dysfunction and insulin resistance in the pathogenesis of type 2 diabetes mellitus. J. Clin. Invest. *104*, 787–794.

Williams, M.D., Joglekar, M.V., Hardikar, A.A., and Wong, W.K.M. (2020). Directed differentiation into insulin-producing cells using microRNA manipulation. Open Med. (Wars) *15*, 567–570.

Wong, W., Farr, R., Joglekar, M., Januszewski, A., and Hardikar, A. (2015). Probe-based real-time PCR approaches for quantitative measurement of microRNAs. J. Vis. Exp. *14*, 52586.

Wong, W.K., Jiang, G., Sorensen, A.E., Chew, Y.V., Lee-Maynard, C., Liuwantara, D., Williams, L., O'connell, P.J., Dalgaard, L.T., Ma, R.C., et al. (2019). The long noncoding RNA MALAT1 predicts human pancreatic islet isolation quality. JCI Insight *5*, e129299.

Wong, W.K.M., Sorensen, A.E., Joglekar, M.V., Hardikar, A.A., and Dalgaard, L.T. (2018). Non-coding RNA in pancreas and beta-cell development. Noncoding RNA *4*, 41.

Xu, G., Chen, J., Jing, G., Grayson, T.B., and Shalev, A. (2016). miR-204 targets PERK and regulates UPR signaling and beta-cell apoptosis. Mol. Endocrinol. *30*, 917–924.

Xu, G., Chen, J., Jing, G., and Shalev, A. (2013). Thioredoxin-interacting protein regulates insulin transcription through microRNA-204. Nat. Med. *19*, 1141–1146.

iScience, Volume *24*

Supplemental information

Machine learning workflows identify a microRNA signature of insulin transcription in human tissues

Wilson K.M. Wong, Mugdha V. Joglekar, Vijit Saini, Guozhi Jiang, Charlotte X. Dong, Alissa Chaitarvornkit, Grzegorz J. Maciag, Dario Gerace, Ryan J. Farr, Sarang N. Satoor, Subhshri Sahu, Tejaswini Sharangdhar, Asma S. Ahmed, Yi Vee Chew, David Liuwantara, Benjamin Heng, Chai K. Lim, Julie Hunter, Andrzej S. Januszewski, Anja E. Sørensen, Ammira S.A. Akil, Jennifer R. Gamble, Thomas Loudovaris, Thomas W. Kay, Helen E. Thomas, Philip J. O'Connell, Gilles J. Guillemin, David Martin, Ann M. Simpson, Wayne J. Hawthorne, Louise T. Dalgaard, Ronald C.W. Ma, and Anandwardhan A. Hardikar

Supplementary Figure 1. Pancreatic gene and microRNA expression analysis, related to Figure 1 and 2

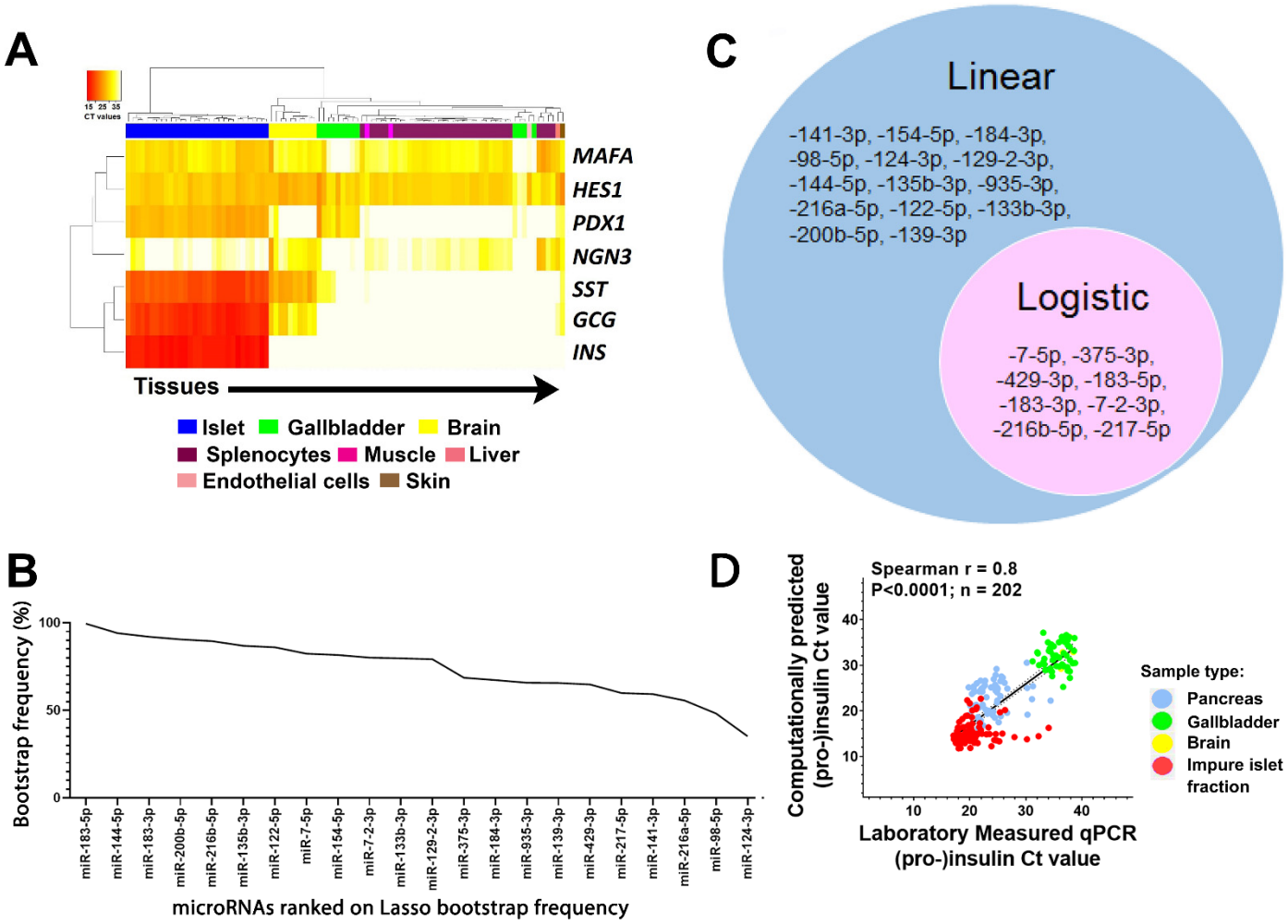
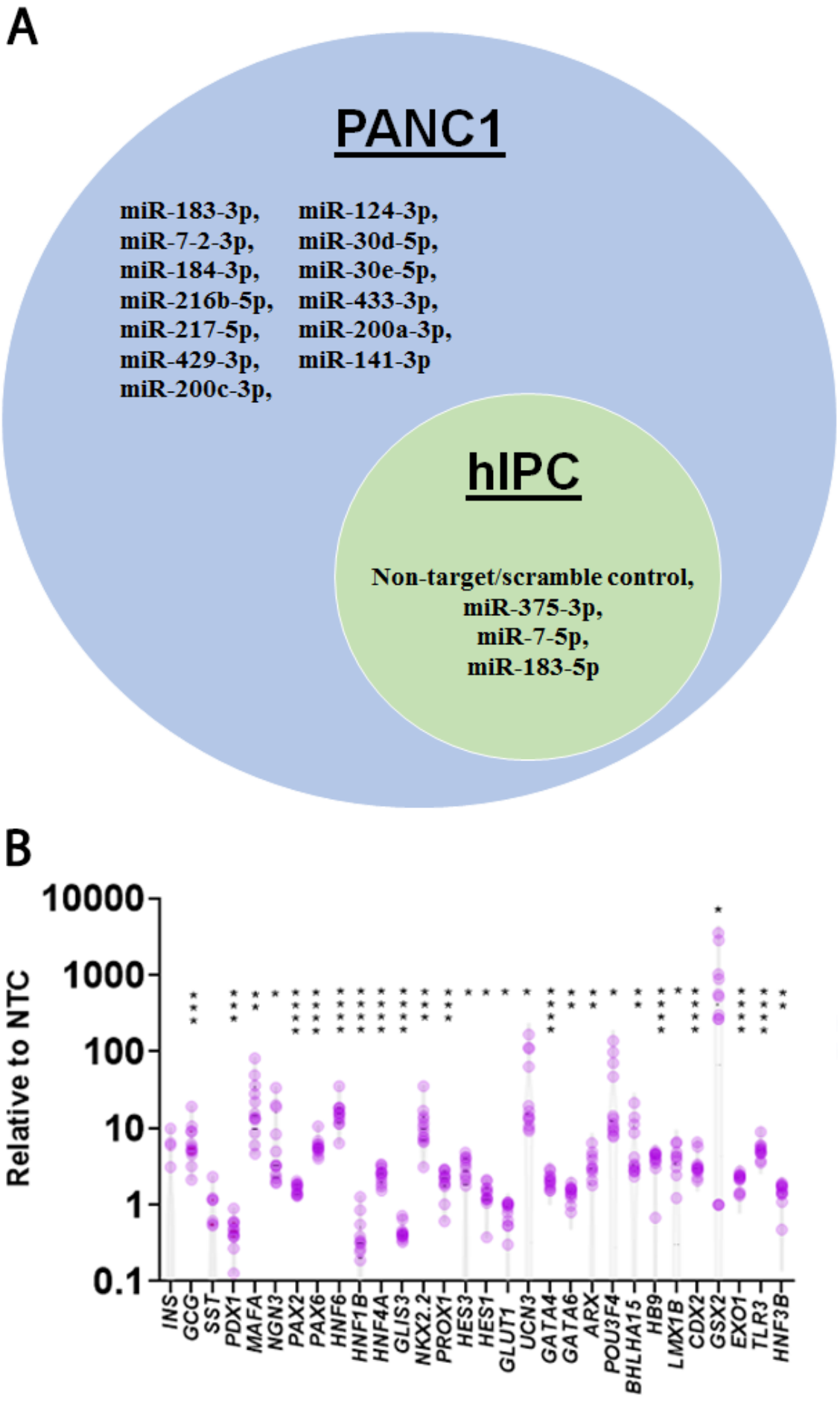


Figure S1. Pancreatic gene and microRNA expression analysis, related to Figure 1 and 2

(A) Unsupervised bidirectional hierarchical cluster analysis for the three pancreatic islet (pro-) hormones and four transcription factors in different human tissue samples from the discovery set (n=92). Lower Ct-values (higher abundance) are presented in shades of red whereas higher Ct-values (low to no expression) appear in the range from yellow to white color. Euclidean distance metric and average linkage were applied to the unsupervised hierarchical cluster analysis. (Ct)-values presented are normalized to 18S rRNA. **(B)** The microRNAs used in prediction of (pro-)insulin transcript (**Figure 2F**) were identified through penalized linear regression workflow and are presented as bootstrap frequency (%) on x-axis using a discovery set of islets (n=30; all insulin-positive) and insulin negative (n=62) samples. **(C)** Euler plot illustrating that the top-eight microRNAs obtained through penalized logistic regression analysis (highlighted in **Figure 1F**) is a subset of the microRNAs obtained from penalized linear regression analysis (**Figure S1B**). (**D**) Correlation plot between the Ct-value for laboratory-measured (pro-)insulin mRNA (x-axis) from the prediction set (n=202) with the computationally predicted (pro-)insulin Ct-value (y-axis) (Spearman r=0.8, p<0.0001). Each dot represents a different sample from the prediction set, while color of the point indicates the tissue type in the correlation plot.

Supplementary Figure 2. MicroRNA-overexpressing cell lines generated through this study, related to Figure 3



Supplementary Figure 2: **MicroRNA-overexpressing cell lines generated through this study, related to Figure 3**

(A) Euler plot of puromycin-selected and doxycycline-regulated microRNA-overexpressing lines generated and used for this study. The specific lines in primary human islet-derived progenitor cells (hIPCs) as well as in human pancreatic duct (PANC1) cells are shown. All of these lines (and associated resources) will be available to academic researchers on reasonable request from lead contact. (B) Transcript abundance of specific transcription factor or (pro-)hormone after the top 8 microRNA lines were mixed and co-cultured in equal proportions (n=11, separate experiments; with each experiment presented as a purple dot). Transcript abundance was calculated relative to the non-targeting/scramble control (NTC, n=4, separate experiments) PANC1 cell line and presented on the Y-axis. *: $p<0.05$, **: $p<0.01$, ***: $p<0.001$, ****: $p<0.0001$; compared to NTC.

Supplementary Figure 3. Inhibition of individual microRNA in human islet cells, related to Figure 6

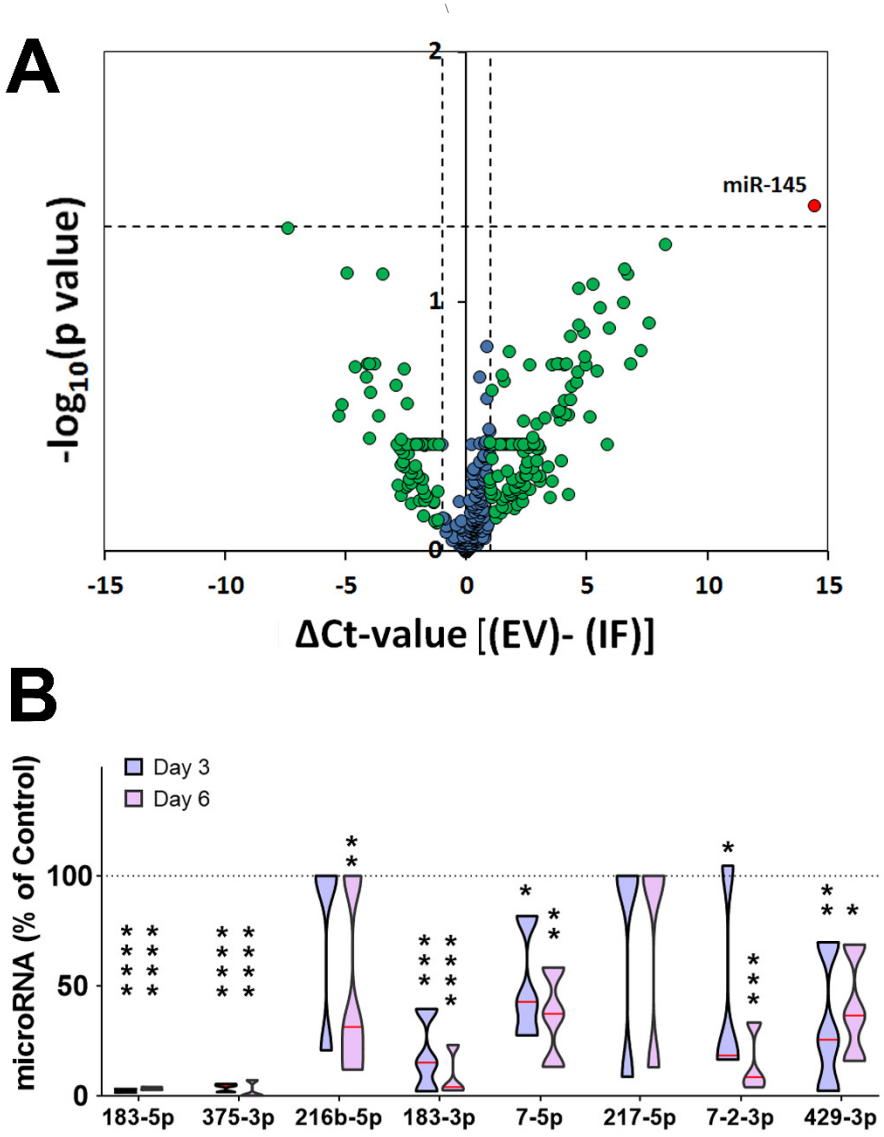


Figure S3: Inhibition of individual microRNA in human islet cells, related to Figure 6

(A) Volcano plot for changes in expression of the 754 (discovery) microRNAs following forced expression of the Furin-cleavable insulin (Ins-Fur/IF-; n=5) vs empty vector (EV-)expressing hIPCs (n=5). The Ct-value differences are depicted on the x-axis and the −log10 p-value (calculated using two-tailed Welch's t-test) is plotted on the y-axis. The dotted horizontal line represents the cut-off of p=0.05, while the dotted vertical lines represent a difference of one Ct-value (2-fold). Each point represents a unique microRNA. None of our bootstrap microRNAs were changed/regulated by forced expression of insulin **(B)** Abundance of the microRNAs following knock-down with LNA power inhibitors in human islet cells (n=3) on day 3 and 6 post-incubation. The horizontal red line represents the median. Polygons represent the density of distribution of the data and extend to min/max values. The y-axis presents microRNA abundance as % of scramble controls (n=3). Statistical significance calculated using two-way ANOVA compared to the control) are presented where *: p<0.05, **: p<0.01, ***: p<0.001, ****: p<0.0001.

Supplementary Figure 4. Insulin-associated microRNA are significantly downregulated in type 2 diabetes, related to Figure 6

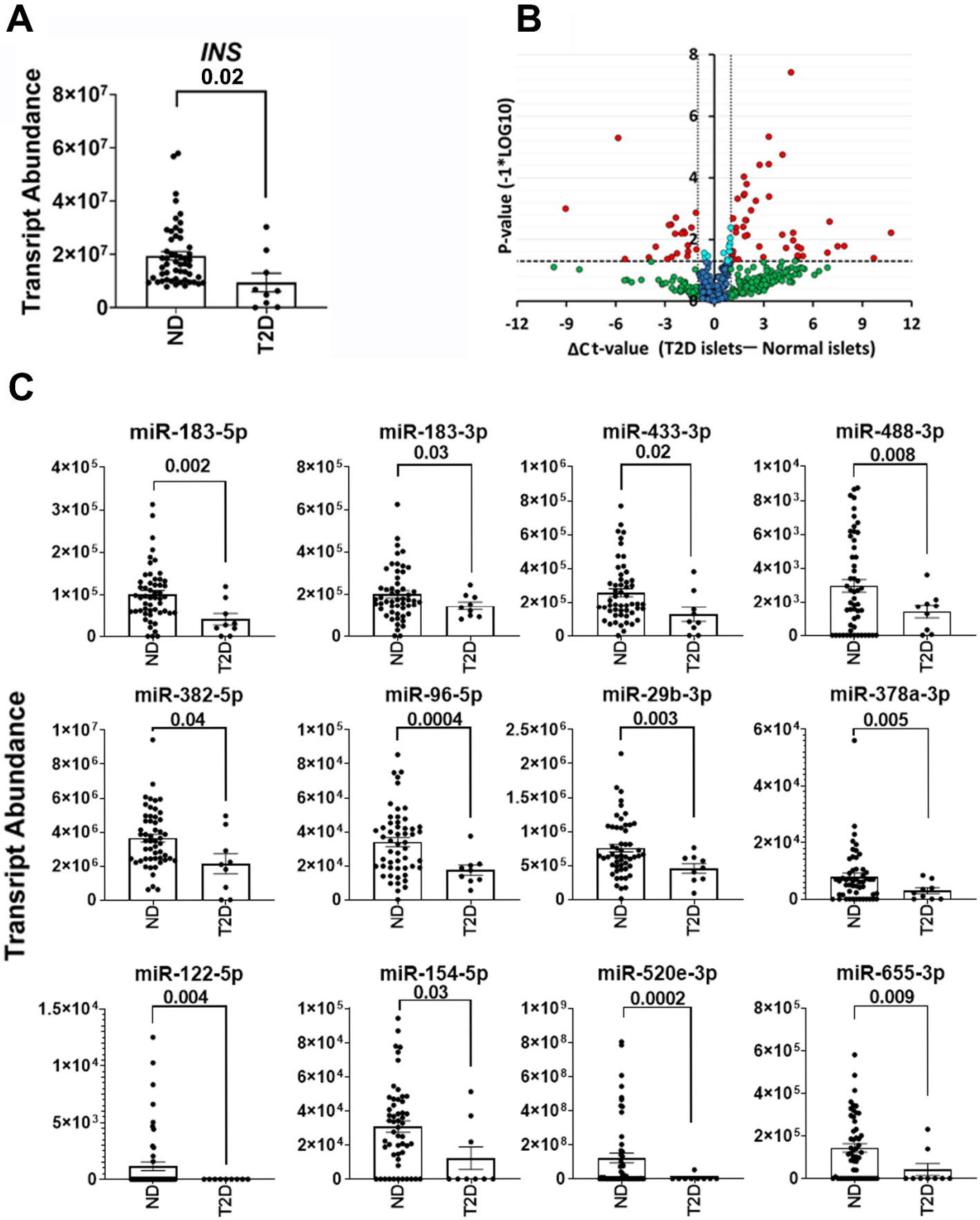


Figure S4: Insulin-associated microRNA are significantly downregulated in type 2 diabetes, related to Figure 6

(**A**) Real-time TaqMan® qPCR data for (pro-)insulin gene transcript abundance in human islets (n=53) from individuals without diabetes vs T2D islets (n=9). Cycle threshold values were normalized to 18S rRNA and then converted to transcript abundance as described(Hardikar et al., 2014). Data are plotted as mean±SEM and analyzed using a two-tailed Welch's t-test to present the exact p-value. (**B**) Volcano plot comparing significantly differentially-expressed microRNAs between T2D islets (n=9) and islets from donors without diabetes (n=53) for all the 754 (discovery) microRNAs. The Ct-value differences are on the X-axis and the −log10 p-value (calculated using two-tailed Welch's t-test) is on the Y-axis. The dashed horizontal line represents the significant p-value=0.05 while the vertical dashed lines represent a difference of 2-fold (1 Ct value). Each point represents a unique microRNA. Significantly different microRNAs are colored in red. (**C**) Several of the microRNAs identified through the penalized logistic regression bootstrap analysis (**Figure 1F**) in islets from donors without diabetes and T2D islets. Global-normalized microRNA data are converted to transcript abundance as described(Hardikar et al., 2014). The mean±SEM is presented for each plot. Data were analyzed using a two-tailed Welch's t-test and exact p-values are provided.

Supplementary Figure 5. microRNA target and KEGG Pathway analysis, related to Figure 6

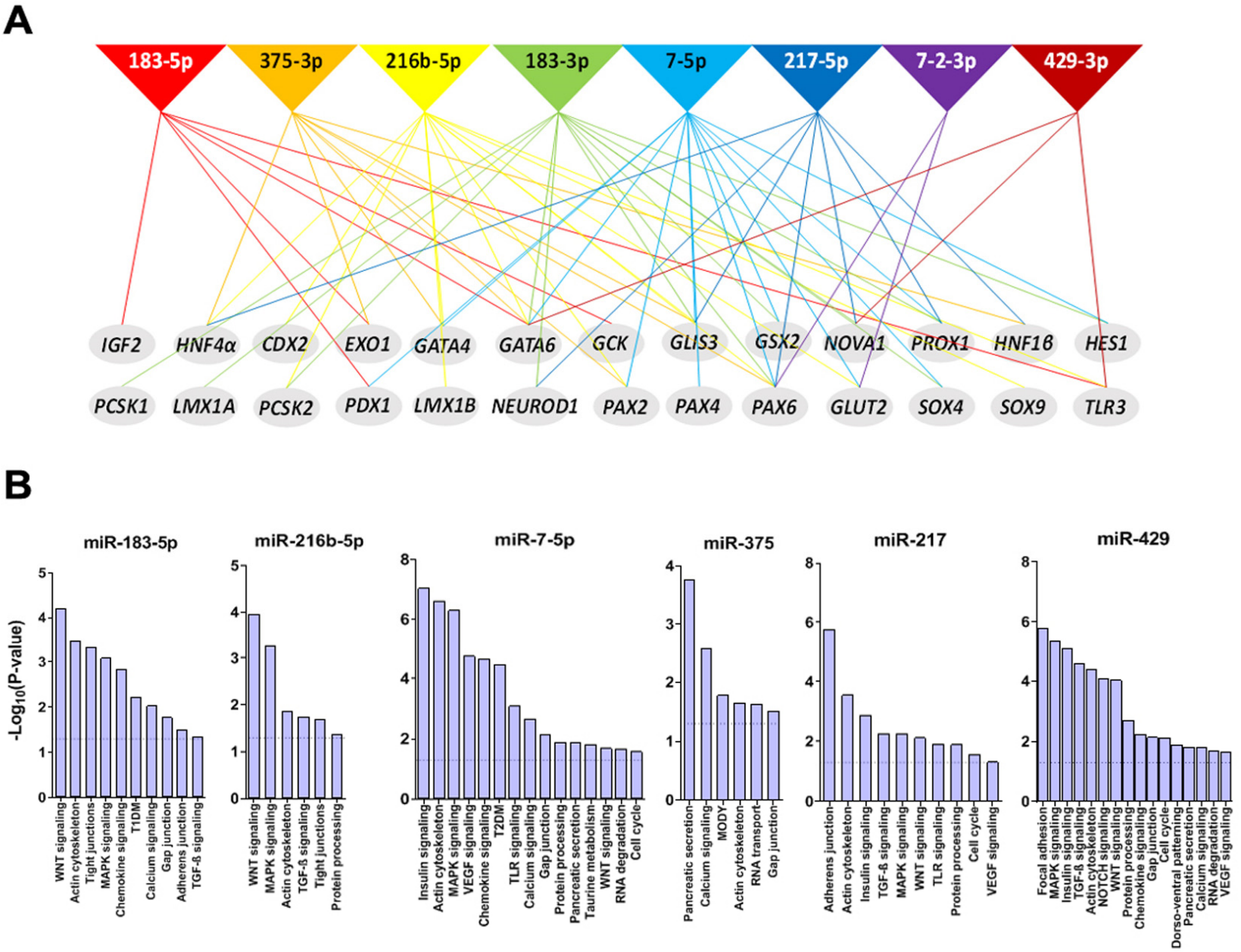


Figure S5: microRNA target and KEGG Pathway analysis, related to Figure 6

(**A**) The top eight microRNAs targeting pancreatic genes were assessed using TargetScan (http://www.targetscan.org/vert_72/) and interactions between each of the top-eight microRNAs and pancreatic genes (selected for this study) are demonstrated as connecting color-coded lines. **(B)** The most significant and relevant KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways associated with key bootstrap microRNAs are presented. The y-axis represents −log10 p-value, the dashed horizontal line represents the significant p-value=0.05 and relevant pathways are provided on the x-axis.

Supplementary Table 1. Human cell/tissue samples assessed in this study, related to Figure 1

Sample type	Discovery set	Validation set	Prediction set	T2D samples
Brain	10(Insulin$^-$)		4 (Insulin$^+$)	
Endothelial cell	1(Insulin$^-$)	14 (Insulin$^-$)		
Gallbladder	13(Insulin$^-$)	11(Insulin$^-$)	51(Insulin$^+$)	
Islet*	30 (Insulin$^+$)	23 (Insulin$^+$)		
Muscle	2(Insulin$^-$)	2 (Insulin$^-$)		
Liver	1 (Insulin$^-$)			
Skin	1 (Insulin$^-$)			
Spleen	34 (Insulin$^-$)			
Impure islet fraction			85 (Insulin$^+$)	
Pancreas*			62 (Insulin$^+$)	
T2D islet				9 (Insulin$^+$)
Total (n)	92	50	202	9
Analysis method for each set	Penalized regression	ROC curve analysis	Correlation	T2D vs ND
Presented in Figure	1	2E	2F and S1D	S4

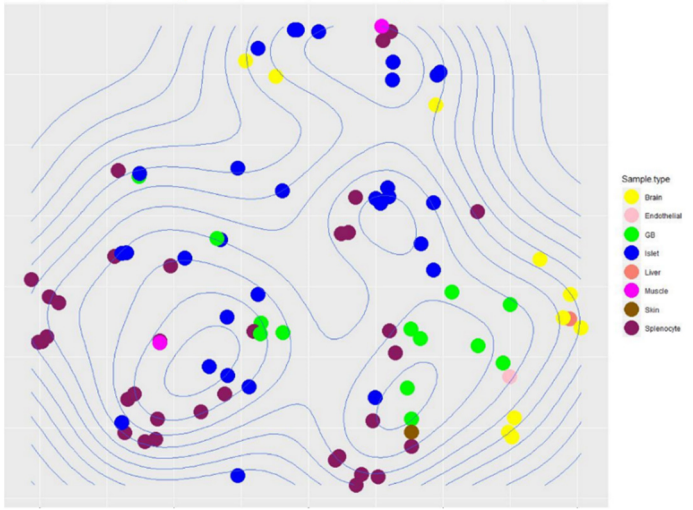
Grand total 353

Supplementary Table 1. Human cell/tissue samples assessed in this study, related to Figure 1.
Different cell/tissue types with their numbers used under the discovery, validation and prediction studies are presented herein. Islets from nine T2D donors were also assessed in a separate set. Impure islet fractions refer to human islet preparations that had a (pro-)insulin Ct-value >16.8. Gallbladder and brain samples show significant heterogeneity in (pro-)insulin transcript as measured by qPCR. The gallbladder and brain samples are therefore sub-classified into "Insulin$^+$" (Ct-value <39) and "Insulin$^-$" (Ct-value $\geq$39) samples. Discovery and validation studies were based on a set of 142 (92 and 50 respectively) tissue samples available to the group and randomly assigned into the discovery and validation set. The discovery set (n=92) was used to identify the key microRNAs associated with insulin transcription via machine learning (penalized regression and bootstrap analysis) as presented in **Figure 1**. The validation set (n=50) was used to assess the ability of the top 8 microRNAs associated with insulin transcript in separating human islets and insulin-negative samples using ROC curve analysis as presented in **Figure 2E**. The prediction set (n=202) was used to assess the correlation between the computationally predicted (pro-)insulin Ct values to those of the laboratory-measured (pro-)insulin Ct values as presented in **Figure 2F and S1D**. The T2D islets (n=9) were presented in **Figure S4**.
*Islets and pancreas from donors without diabetes (ND).

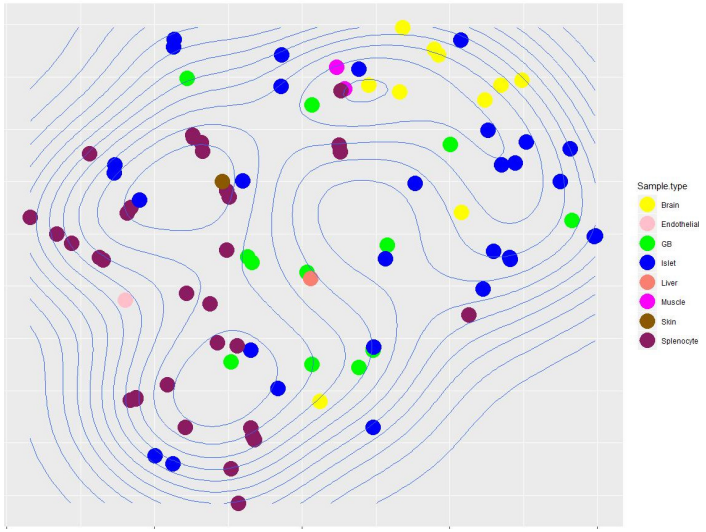
Supplementary Table 3. MicroRNAs used in t-SNE analysis, related to Figure 1

Top eight miRNAs in penalized logistic regression bootstrap	Total # of samples (and %) present	Random eight miRNAs	Total # of samples (and %) present
miR-183-5p	37 (40)	miR-516b-3p	38 (41)
miR-375-3p	77 (84)	miR-601-5p	77 (84)
miR-216b-5p	50 (54)	miR-1290-3p	50 (54)
miR-183-3p	39 (42)	miR-624-5p	39 (42)
miR-7-5p	78 (85)	miR-664a-3p	78 (85)
miR-217-5p	38 (41)	miR-17-3p	38 (41)
miR-7-2-3p	43 (47)	miR-638-5p	43 (47)
miR-429-3p	72 (78)	miR-511-5p	72 (78)
miR-183-5p	37 (40)	miR-651-5p	36 (39)
miR-375-3p	77 (84)	miR-629-5p	78 (85)
miR-216b-5p	50 (54)	miR-34c-5p	50 (54)
miR-183-3p	39 (42)	miR-92a-1-5p	39 (42)
miR-7-5p	78 (85)	miR-26b*	78 (85)
miR-217-5p	38 (41)	miR-888-5p	38 (41)
miR-7-2-3p	43 (47)	miR-34b-3p	43 (47)
miR-429-3p	72 (78)	miR-146b-3p	72 (78)

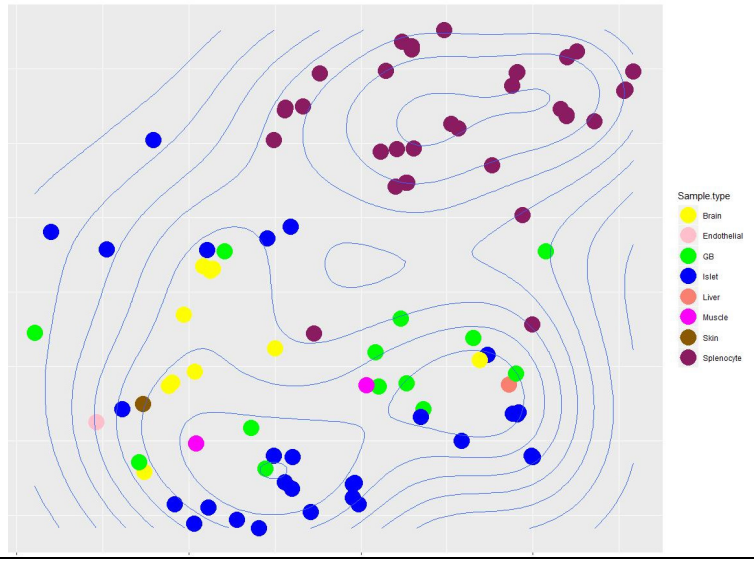
Set 1 (shown in Figure 1H)



Set 2



				Set 3
miR-183-5p	37 (40)	miR-550a-5p	36 (39)	
miR-375-3p	77 (84)	miR-192-3p	76 (83)	
miR-216b-5p	50 (54)	miR-224	50 (54)	
miR-183-3p	39 (42)	miR-24-2-5p	40 (43)	
miR-7-5p	78 (85)	miR-539-5p	77 (84)	
miR-217-5p	38 (41)	miR-100-3p	39 (42)	
miR-7-2-3p	43 (47)	miR-627-5p	43 (47)	
miR-429-3p	72 (78)	miR-495-3p	72 (78)	



Supplementary Table 3. MicroRNAs used in t-SNE analysis, related to Figure 1

Three different sets of eight random miRNAs (shown in column 3) were selected with a similar detectability (present) to the top-eight microRNAs obtained from the logistic regression analyses (**Figure 1F** and shown here in the first column). The similar detectability/presence between the corresponding top-eight microRNA to the random eight microRNA (within the discovery set of samples; n=92) are shown in columns 2 and 4 of this table respectively. The t-SNE plot (on the discovery set; n=92) displayed on the right is generated from using the corresponding random set of eight microRNAs shown on the left. The t-SNE plot using the top eight microRNAs for the discovery set samples (n=92) is presented in **Figure 1G.** The above table presents the t-SNE plots for three sets of random eight microRNAs with levels/distribution matched to the top eight microRNAs identified through our analysis.

Supplementary Table 4. Correlation between microRNA and (pro-)insulin expression, related to Figure 2

vs INS	Pearson r	Pearson p-value (two-tailed)	Pearson 95% confidence interval
miR-183-5p	0.6	<0.0001	0.5 to 0.7
miR-375-3p	0.7	<0.0001	0.6 to 0.8
miR-216b-5p	0.6	<0.0001	0.4 to 0.7
miR-183-3p	0.6	<0.0001	0.5 to 0.7
miR-7-5p	0.7	<0.0001	0.6 to 0.8
miR-217-5p	0.5	<0.0001	0.3 to 0.6
miR-7-2-3p	0.7	<0.0001	0.6 to 0.8
miR-429-3p	0.7	<0.0001	0.5 to 0.7

Supplementary Table 4. Correlation between microRNA and (pro-)insulin expression, related to Figure 2

Pearson correlation analysis of the individual microRNA (belonging to the top 8 microRNAs, highlighted in **Figure 1F**) with the respective (pro-)insulin transcript expression from islets (n=30), gallbladder (n=51), brain (n=4) and spleen (n=34; as shown in **Figure 2D**). Pearson's r and p-value (two-tailed) along with 95% confidence interval are provided for each microRNA.

Supplementary Table 5. Gene assay IDs used for qPCR, related to Figure 2

	Assay ID	Gene symbol	Gene name	Dye	Stock
1	Hs03003631_g1	*18S*	Eukaryotic 18S rRNA	VIC-MGB	60x*
2	Hs02741908_m1	*INS*	Insulin	FAM-MGB	20x
3	Hs00355773_m1	*INS*	Insulin	FAM-MGB	20x
4	Hs00356618_m1	*INS*	Insulin	FAM-MGB	20x
5	Hs00174967_m1	*GCG*	Glucagon	FAM-MGB	20x
6	Hs00174949_m1	*SST*	Somatostatin	FAM-MGB	20x
7	Hs01651425_s1	*MAFA*	V-Maf Avian Musculoaponeurotic Fibrosarcoma Oncogene Homolog A	FAM-MGB	20x
8	Hs00360700_g1	*NGN3*	Neurogenin 3	FAM-MGB	20x
9	Hs00172878_m1	*HES1*	Hairy and Enhancer of Split 1 (Hes Family BHLH Transcription Factor 1)	FAM-MGB	20x
10	Hs00236830_m1	*PDX1*	Pancreatic And Duodenal Homeobox 1	FAM-MGB	20x

Supplementary Table 5. Gene assay IDs used for qPCR, related to Figure 2

List of TaqMan® primer/probe assays selected for real-time qPCR on ViiA7 platform. *A one in three dilution was performed to the stock 18S (60x) to match the probe-primer concentration of other assays.

Supplementary Table 6. Pancreatic gene and transcription factors used on customized OpenArray™ panel, related to Figure 3 and 5

	Assay ID	Gene Symbol	Gene Name
1	Hs00173014_m1	*PAX4*	Paired box 4
2	Hs00236830_m1	*PDX1*	Pancreatic And Duodenal Homeobox 1
3	Hs00355773_m1	*INS*	Insulin
4	Hs01031536_m1	*GCG*	Glucagon
5	Hs00356144_m1	*SST*	Somatostatin
6	Hs01875204_s1	*NEUROG3*	Neurogenin 3 (referred herein as *NGN3*)
7	Hs01651425_s1	*MAFA*	V-Maf Avian Musculoaponeurotic Fibrosarcoma Oncogene Homolog A
8	Hs00271378_s1	*MAFB*	V-Maf Avian Musculoaponeurotic Fibrosarcoma Oncogene Homolog B
9	Hs00359592_m1	*NOVA1*	Neuro-Oncological Ventral Antigen 1
10	Hs00240858_m1	*PAX2*	Paired Box 2
11	Hs03003631_g1	*18S*	Eukaryotic 18S rRNA
12	Hs00240871_m1	*PAX6*	Paired Box 6
13	Hs00268388_s1	*SOX4*	SRY-Box Transcription Factor 4
14	Hs01001343_g1	*SOX9*	SRY-Box Transcription Factor 9
15	Hs00603586_g1	*PTF1A*	Pancreas Associated Transcription Factor 1a
16	Hs00413554_m1	*HNF6 (ONECUT1)*	Hepatocyte Nuclear Factor 6 (One Cut Homeobox 1)
17	Hs01001602_m1	*HNF1β (TCF2)*	Hepatocyte Nuclear Factor 1-Beta (Transcription Factor 2)
18	Hs00167041_m1	*HNF1α (TCF1)*	Hepatocyte Nuclear Factor 1-Alpha (Transcription Factor 1)
19	Hs00230853_m1	*HNF4α*	Hepatocyte Nuclear Factor 4-Alpha
20	Hs00541450_m1	*GLIS3*	GLIS Family Zinc Finger 3
21	Hs00232355_m1	*NKX6.1*	NK6 Transcription Factor Related, Locus 1
22	Hs00356618_m1	*INS*	Insulin
23	Hs00159616_m1	*NKX2.2*	NK2 Transcription Factor Related, Locus 2
24	Hs00896294_m1	*PROX1*	Prospero-Related Homeobox 1
25	Hs01367669_g1	*HES3*	Hairy And Enhancer Of Split 3 (Hes Family BHLH Transcription Factor 3)
26	Hs00172878_m1	*HES1*	Hairy And Enhancer Of Split 1 (Hes Family BHLH Transcription Factor 1)
27	Hs01922995_s1	*NEUROD1*	Neuronal Differentiation 1
28	Hs00892681_m1	*GLUT1 (SLC2A1)*	Solute Carrier Family 2 (Facilitated Glucose Transporter), Member 1
29	Hs01096908_m1	*GLUT2 (SLC2A2)*	Solute Carrier Family 2 (Facilitated Glucose Transporter), Member 2
30	Hs01564555_m1	*GCK*	Glucokinase
31	Hs00846499_s1	*UCN3*	Urocortin 3
32	Hs00158126_m1	*ISL1*	ISL LIM Homeobox 1
33	Hs00171403_m1	*GATA4*	GATA Binding Protein 4

34	Hs00232018_m1	*GATA6*	GATA Binding Protein 6
35	Hs00292465_m1	*ARX*	Aristaless Related Homeobox
36	Hs00355773_m1	*INS*	Insulin
37	Hs01005963_m1	*IGF2*	Insulin Like Growth Factor 2
38	Hs00264887_s1	*POU3F4*	POU Class 3 Homeobox 4
39	Hs02758991_g1	*GAPDH*	Glyceraldehyde-3-Phosphate Dehydrogenase
40	Hs00703572_s1	*BHLHA15 (MIST1)*	Basic Helix-Loop-Helix Family Member A15
41	Hs00907365_m1	*HB9 (MNX1)*	Homeobox HB9 (Motor Neuron And Pancreas Homeobox 1)
42	Hs00158750_m1	*LMX1.2 (LMX1B)*	LIM Homeobox Transcription Factor 1 Beta
43	Hs00892663_m1	*LMX1.1 (LMX1A)*	LIM Homeobox Transcription Factor 1 Alpha
44	Hs01078080_m1	*CDX2*	Caudal Type Homeobox 2
45	Hs00793699_g1	*GSX1*	GS Homeobox 1
46	Hs02741908_m1	*INS*	Insulin
47	Hs00370195_m1	*GSX2*	GS Homeobox 2
48	Hs01116195_m1	*EXO1*	Exonuclease 1
49	Hs00170171_m1	*REG3A*	Regenerating Family Member 3 Alpha
50	Hs01551078_m1	*TLR3*	Toll Like Receptor 3
51	Hs00358111_g1	*PPY*	Pancreatic Polypeptide
52	Hs00230829_m1	*AIRE*	Autoimmune Regulator
53	Hs01074053_m1	*GHRL*	Ghrelin And Obestatin Prepropeptide
54	Hs01026107_m1	*PCSK1*	Proprotein Convertase Subtilisin/Kexin Type 1
55	Hs00159922_m1	*PCSK2*	Proprotein Convertase Subtilisin/Kexin type 2
56	Hs00232764_m1	*HNF3β (FOXA2)*	Hepatocyte Nuclear Factor 3-Beta (Forkhead Box A2)

Supplementary Table 6. Pancreatic gene and transcription factors used on customized OpenArray™ panel, related to Figure 3 and 5

List of the TaqMan® primer/probe gene expression assays for real-time qPCR on the OpenArray™ platform. All assays use FAM-MGB dye at 20x stock concentration. Three different assays for (pro-)insulin present on the OpenArray™ panel are highlighted in yellow of which one assay (Hs00355773_m1) was placed on two (geographically) separate areas of the panels (well #3 and #36) as quality control to assess for any differences in the measurement of the same gene assay across different regions of the plate. Matching samples across different plates confirmed no significant variation in the lots/batches of custom plates. Housekeeping genes (18S and *GAPDH*) are highlighted in grey.

Transparent Methods

Sample collection and storage

All tissue samples were obtained through human research ethics committee (HREC) approvals X16-0289 (previously X12-0176) and the HREC/12/RPAH/282 as well as MQ5201300330 protocols. Human pancreas from individuals without diabetes, impure and pure islet samples (from donors without diabetes and T2D donors) were obtained as part of the research consented tissues through the National Islet Transplantation Program (at Westmead Hospital, Sydney and the St Vincent's Institute, Melbourne, Australia). Human spleen samples were obtained from collaborating investigators at the Westmead Hospital, Sydney, Australia. Human gallbladder and muscle samples were obtained as surgical waste tissues generated following cholecystectomies or other Gastrointestinal (GI) surgeries involving non-cancerous tissue through the surgical teams (at Strathfield Private Hospital and The Royal Prince Alfred Hospital, Australia). Human gallbladder epithelial cells were separated in the lab by scraping off the epithelial cells with a disposable scalpel blade after washing off the bile under aseptic conditions. Cells were resuspended in Phosphate Buffered Saline (PBS), visualized under the microscope and then stored immediately as a dry pellet for subsequent RNA isolation. The human brain, liver, and skin tissue samples were the only developing tissue samples (around 16-20 weeks gestation age) acquired from collaborating investigators at Macquarie University, Australia. All tissue samples were obtained in the transport medium, washed thrice with generous amounts of dPBS (1x) (Gibco, Thermo Fisher Scientific, Waltham, MA) and snap-frozen immediately as a dry pellet. Human Umbilical Vein Endothelial Cells (HUVECs) were obtained through collaborators at the Centenary Institute, Sydney, Australia. The umbilical cord was washed by passing through generous volumes of sterile dPBS (1x) until no traces of blood were detected in the wash through. Endothelial cells were isolated following collagenase digestion and washed in culture media. HUVECs were cultured in $25cm^2$ (T25) flasks coated with 1ml of gelatine (Sigma-Aldrich, St Louis, MO) in HUVEC medium (M199 with Earle's Salts, 20mM HEPES, 20% fetal calf serum, sodium bicarbonate, 2mM glutamine, 1% non-essential amino acids (NEAA), 1mM sodium pyruvate, penicillin and gentamicin (all from Invitrogen, Thermo Fisher Scientific, Waltham, MA). Endothelial cells were maintained at 37°C, in 5% CO_2 and used for RNA isolation within 3-4 days. All samples were stored at -80°C until RNA isolation.

RNA isolation and Quantification

Total RNA isolation from samples was carried out using the TRIzol® (Thermo Fisher Scientific, Waltham, MA). Tissue samples were first homogenized on ice in 200µl of TRIzol® reagent and then the volume was made up to 1ml. Total RNA was isolated from cell and tissue samples using the standard manufacturer's protocol with minor modification as described elsewhere(Joglekar and Hardikar, 2012). A small subset of human islet samples was used to isolate microRNA and mRNA using the mirVana miRNA isolation kit (Thermo Fisher Scientific, Waltham, MA) following the manufacturer's instruction. RNA quality and quantity were assessed using the Nanodrop™ spectrophotometric platform, where all samples (**Table S1**) were found to have an average 260/280 of 1.84. Small RNA and miRNA content were measured using the 2100 Bioanalyzer Small RNA kit (Agilent Technologies, Santa Clara, CA) as per manufacturer's instructions.

Synthesis of cDNA and TaqMan® real-time qPCR

Synthesis of cDNA (from total RNA or enriched mRNA) was carried out using the High Capacity cDNA reverse transcription kit (Thermo Fisher Scientific, Waltham, MA). TaqMan® real-time qPCR was performed in 5µl reactions using 96-well plates with 33.3ng input cDNA/reaction and TaqMan® Fast Universal PCR Master Mix (Thermo Fisher Scientific, Waltham, MA). Ten TaqMan® primer/probe gene expression assays were selected (**Table S5**) to quantify transcript abundance on the ViiA7 platform (Thermo Fisher Scientific, Waltham, MA). The amplification curve in the linear region was set at a threshold of 0.1 for all samples. Our samples (**Table S1**) had an average 18S rRNA cycle threshold (Ct) value of 9.06. Gene expression data were normalized to 18S. The cut-off of normalized (pro-)insulin Ct-value$\leq$16.8 was chosen to identify pure islet preparations while Ct-value >16.8 was used for islet preparations with impurities. For all other tissues, when Ct-value for normalized (pro-)insulin was < 39, they were considered as insulin-positive and if $\geq$39, then insulin-negative. Ct-value of 39 is the limit of detectability measured using a gold-standard TaqMan-based RT-qPCR system (ViiA-7, ThermoFisher, USA) (Hardikar et al., 2014).

Nanofluidics-based TaqMan® real-time PCR on 754 validated human miRNAs

MicroRNA transcript abundance was quantified using the TaqMan® OpenArray™ human microRNA Panels on the QuantStudio™ 12K Flex platform (Thermo Fisher Scientific,

Waltham, MA). Reverse transcription and pre-amplification were carried out using the megaplex Reverse Transcription (RT)/Pre-amplification (PA) primer pools as per manufacturer's protocol described elsewhere(Wong et al., 2015). Briefly, each sample (100ng input), underwent 12 pre-amp cycles, then diluted 1:40 in 0.1x TE (pH 8.0), combined with TaqMan® OpenArray™ PCR Master Mix, loaded onto the TaqMan® OpenArray™ Human MicroRNA Panel using the robotic AccuFill™ system (Thermo Fisher Scientific, Waltham, MA) and qPCR was completed using the QuantStudio™ 12K Flex. The microRNA off-the-shelf panel includes 754 known/validated human miRNA assays. All acquired data was uploaded, and global normalization protocol was applied through the Thermo Fisher Cloud data analysis workflow (Thermo Fisher Scientific, Waltham, MA). Global normalization method has been described previously (Mestdagh et al., 2009) that identifies the assays common to all samples and uses their geometric mean value as the normalization factor for each sample. PCR results below the amplification score of 1.24 were classified as having no amplification and were considered undetectable for further analyses.

Digital droplet PCR

Digital droplet PCR (ddPCR) was carried out on a Bio-Rad QX200™ Droplet Digital™ PCR System (Bio-Rad, Hercules, CA) with an automated droplet generator as per manufacturer's instructions. The Insulin TaqMan® primer/probe assay (Hs02741908_m1) (Thermo Fisher Scientific, Waltham, MA) was used to measure (pro-)insulin expression in human islet, gallbladder, brain, endothelial cell, muscle and spleen samples on ddPCR. For each human sample – a total of 125ng cDNA input was used per reaction, except for human islet samples that had to be diluted to avoid signal saturation on the ddPCR system. For each human islet sample, a 0.125ng input of cDNA was used per reaction. Nuclease-free water was used as a no-template control to set the threshold. Plate set up and calculations were performed as described before(Maynard et al., 2019). Some data were obtained using QuantStudio™ 3D Digital PCR system (Thermo Fisher Scientific, Waltham, MA) with the same insulin primer/probe assay.

Cell culture

PANC1 cells were maintained in serum-containing medium (SCM) prepared by mixing high glucose (4.5g/L) DMEM with 1% GlutaMAX™, 10% fetal bovine serum, 100U/mL penicillin and 100µg/mL streptomycin. Cells were maintained in an incubator at 37°C and with humidified 5% CO_2 in air. PANC1 cells were passaged 1:3 when confluent.

Around 5000 freshly isolated human islets were cultured in 75cm^2 tissue culture-treated flasks with 10ml CMRL-1066 medium containing 1x GlutaMAX™, 100U/mL penicillin and 100µg/mL streptomycin and 10%(v/v) fetal bovine serum (all media and reagents from Thermo Fisher Scientific, Waltham, MA) and 10ng/mL epithelial growth factor (Sigma-Aldrich, St Louis, MO). Cells were maintained in an incubator at 37°C and with humidified 5% CO_2 in air and passaged 1:2 when confluent. Human islet cells attached, migrated and then proliferated for several passages *in vitro* to generate human islet-derived progenitor cells (hIPCs) that typically exhibit mesenchymal characteristics and do not contain any (pro-)insulin transcripts(Joglekar et al., 2009, Gershengorn et al., 2004).

Human islet cells at around day 3-12 in culture ((pro-)insulin transcript positive) were used for microRNA LNA inhibitor experiments, whereas human islets expanded/grown for several passages ((pro-)insulin transcript negative; hIPCs) were used for microRNA and insulin overexpression experiments as described below.

Generation of microRNA-overexpressing PANC1 and hIPCs lines

Third-generation doxycycline-inducible and puromycin-resistant SMARTvector™ shRNA lentiviral vectors with GFP or RFP reporters and transcribing the desired mature miRNA sequences (Dharmacon, Lafayette, CO) were used. PANC1 and hIPCs were seeded in 12-well culture plates and lentiviral vectors along with 8µg/mL polybrene (Sigma-Aldrich, St Louis, MO) were added. The multiplicity of infection (MOI) of 0.2 was used to obtain most of the transduced cells with a single integrated copy. After 24 hours of incubation, the media was changed. After 3 days, the cells were grown in medium containing optimal puromycin (Thermo Fisher Scientific, Waltham, MA) concentration of 8µg/mL for PANC1 or 2µg/mL for hIPCs, for another 7 days, to eliminate any remaining untransduced cells. Transduced cells were grown in the presence of doxycycline (Sigma-Aldrich, St Louis, MO) thereafter (1000ng/mL for PANC1 cells and 200ng/mL for hIPCs). One non-target/scramble control and 16 different microRNA-overexpressing PANC1 cell lines were generated (**Figure S2A**). At least six separate experiments were performed by three different wet-lab researchers for each of the 17 PANC1 lines generated. For hIPCs, one non-target/scramble control and three different microRNA-overexpressing lines were generated with at least three different biological replicates/ pancreas donor islets (**Figure S2A**).

Forced expression of insulin in hIPCs

Insulin transcript-negative hIPCs (from n=5 different organ donors) that typically exhibit mesenchymal characteristics(Joglekar et al., 2009, Gershengorn et al., 2004) were generated and transduced with HMD-INS-FUR(Ren et al., 2007) vector to overexpress insulin. FACS analysis of eGFP expression in hIPCs transduced with the INS-FUR vector (relative to empty vector) was performed on a BD LSR II™ in the Microbial Imaging Facility (UTS, Australia). Flow cytometry data were analyzed using BD FACSDiva™ software (Version 8.0.1). The top 20% eGFP-expressing cells were sorted and either cultured or processed for RNA isolation.

Differentiation of microRNA-overexpressing hIPCs

Day 0 serum-free medium (SFM) was prepared in CMRL by mixing 1% bovine serum albumin (BSA), 1x Insulin-Transferrin-Selenium (ITS-G), 1x GlutaMAX™, 100U/mL penicillin and 100µg/mL streptomycin. Day 4 serum-free medium is a day 0 SFM containing 0.3mM taurine. Day 10 SFM is day 0 SFM containing 0.3mM taurine, 1mM nicotinamide, 100nM exendin-4 (all media and reagents from Thermo Fisher Scientific, Waltham, MA). Three different biological preparations of hIPC lines overexpressing microRNAs and a non-target/scramble control were grown in 75cm^2 flasks in SCM. Cells were washed with 1x PBS, trypsinized using 0.25% trypsin-EDTA and seeded into 24-well suspension plates (Greiner Bio-One, Austria) in day 0 SFM containing 200ng/mL doxycycline. On day 1, media was replaced with day 0 SFM; on day 4 and day 7, media was replaced with day 4 SFM; and on day 10 media was replaced with day 10 SFM. Cells were harvested on day 14 for gene transcript analysis. Experiments were performed with four lines and three different biological preparations and repeated at least two times.

Differentiation of microRNA-overexpressing PANC1 lines

Differentiation was set up as per the schematic in **Figure 5A**, where cells were incubated with 100nM of each miRNA Power LNA inhibitor (Qiagen, Hilden, Germany). After three days, cells were trypsinized and allowed to aggregate in Day 0 SFM on 24-well suspension culture plates. Day 0, Day 4, and Day 10 SFM were prepared as mentioned in hIPCs differentiation methodology above, using low-glucose DMEM (1g/L D-glucose) and Ham's F-12 nutrient mix in a 1:1 ratio (Thermo Fisher Scientific, Waltham, MA) instead of CMRL. In all conditions, Day 4 SFM was added on day 4 and Day 10 SFM was added on day 7, with additional doxycycline added on day 10. Cells were harvested on day 14 for gene transcript analysis. Three to five separate experiments were performed for N, M, and ML conditions (**Figure 5A**).

MicroRNA inhibition in human islets cells

MicroRNA expression was knocked down using miRNA Power LNA inhibitors (Qiagen, Hilden, Germany). Human islet cells (n=3-5 biological replicates) were incubated with 500nM of miRNA Power LNA inhibitor in SCM on a tissue culture plate. Untransfected and negative siRNA (scramble) transfected human islet cells were used as controls. There were no significant differences between the negative siRNA and the untransfected controls. Cell culture medium was not changed and cells were harvested on Day 3 and Day 6.

Static glucose-stimulated insulin secretion (GSIS)

Three conditions of the day 14 differentiated PANC1 cell clusters (N, M, and ML; **Figure 5A**) were washed in 1x dPBS and then with basal glucose medium (2.5mM D-glucose in 0.1% BSA in 1x dPBS). Clusters were incubated in 150µL basal glucose medium for 1h at 37°C, 100µL of the supernatant was then collected and stored at -80°C. Next, 100µL of high glucose medium was added to yield a final concentration of 25mM glucose and incubated for 1h at 37°C. Clusters were gently spun down, and 150µL of supernatant was collected. The GSIS cell pellets were stored in acidified ethanol (0.18M HCl in 70% ethanol) for insulin content and protein measurement (for C-peptide normalization) and analysis, whilst the supernatants (devoid of any cells) were frozen at -80°C for further analyses.

C-peptide ELISA

GSIS cell pellets (described above) were sonicated and re-suspended in acid ethanol (0.18M HCl in 70% ethanol). Ultrasensitive C-peptide ELISA kit (Mercodia, Sweden) was used to measure C-peptide from GSIS supernatants (C-peptide released in response to glucose) as well as the GSIS cell pellets (cellular content for C-peptide), using the protocol provided with the kit. Protein content was measured on Qubit (Thermo Scientific, Waltham, MA) following the manufacturer's instructions.

Probe-based microRNA TaqMan® Real-time qPCR

MicroRNAs used to generate overexpressing lines were quantified using ultra-sensitive TaqMan® qPCR chemistry, either on nanofluidic high-throughput QuantStudio^{TM} 12K Flex platform or on the ViiA7 platform (both from Thermo Fisher Scientific, Waltham, MA). Measurement of microRNAs (such as miR-433-3p, miR-200a-3p, miR-30e-5p, miR-141-3p) was carried out on the ViiA7 platform following our previously reported protocol(Wong et al., 2015). Remaining microRNAs were measured on a customized microRNA panel using the low

sample input (LSI) protocol supplied by the manufacturer (Thermo Scientific, Waltham, MA). Briefly, RNA was reverse transcribed, pre-amplified for 16 cycles and then diluted 1 in 20 in 0.1x TE (pH 8.0) before amplification on the QuantStudio^{TM} 12K Flex platform. Data with amplification score <1.24 and Cq confidence interval <0.6 was converted to undetectable and then normalized to ath-miR-159a spike-in added during reverse transcription. Transcript abundance of a microRNA was calculated based on its normalized Ct-value, using the formula "transcript abundance=$2^{(39\text{-Ct value})}$", where '39' is the Cycle threshold (Ct) limit of detection(Hardikar et al., 2014). Relative transcript abundance was calculated as a difference between transcript abundance values of a test sample and control (NTC/scramble lines).

Pancreatic gene measurement using OpenArray™ Custom mRNA Panel

A customized OpenArray™ Human mRNA panel containing 56 selected TaqMan® primer/probe gene expression assays (Thermo Fisher Scientific, Waltham, MA), relevant to pancreas development was designed (**Table S6**). The customized panel was used to quantify mRNA profile of the 56 genes in the PANC1 lines using the robotic QuantStudio^{TM} 12K Flex Real-Time PCR platform. Briefly, cDNA was synthesized using a High Capacity cDNA reverse transcription kit (Thermo Fisher Scientific, Waltham, MA). A total of 120ng cDNA was pre-amplified for 14 cycles, then diluted 1 in 20 in 0.1x TE (pH 8.0) and amplified on the qPCR OpenArray™ platform. Data with amplification score <1.24 and Cq confidence interval <0.6 was converted to undetectable and then normalized to GAPDH. Transcript abundance was calculated as described above.

Pancreatic single-cell (sc)RNA sequencing analysis

Human pancreatic single-cell sequencing data (n=14,890) was extracted from public datasets (GSE84133, GSE85241, E-MTAB-5061, GSE83139, GSE81608). The Panc8 dataset was analyzed via R (version 3.6.1) on R studio version 1.2.5033 using the SeuratData (version 0.2.1), Seurat (3.2.2), ggplot2 (3.3.3) and cowplot (1.1.1) package. The Seurat Panc8 package(Stuart et al., 2019) contains eight different pancreas single-cell RNA-sequencing datasets from across five technologies (inDrop, CEL-Seq1, CEL-Seq2, Smart-Seq2, and Fluidigm C1). To improve the integrity of the data, low read count datasets obtained through the inDrop technology were excluded. Analytical workflow included data preprocessing and feature selection, dimension reduction and identification of "anchor" correspondences between datasets, filtering, scoring, and weighting of anchor correspondences, and data matrix correction, or data transfer across experiments as described elsewhere(Stuart et al., 2019).

Immunostaining

Freshly isolated human islets, gallbladder epithelial cells and brain neurospheres (generated *in vitro* and grown over coverslips) were fixed in 4% fresh paraformaldehyde, permeabilized using 0.1% Triton X-100, blocked in 4% normal donkey serum (all from Sigma-Aldrich, St Louis, MO), and then incubated overnight with primary antibodies at 4°C. Cells were thoroughly washed with 1x dPBS and incubated with secondary antibodies at 37°C for 1h. Both primary and secondary antibodies were diluted in 4% normal donkey serum (blocking buffer). Cells were washed again with 1x dPBS and mounted in Vectashield™ mountant (Vector laboratories, Burlingame, CA) containing Hoechst 33342 (Thermo Fisher Scientific, Waltham, MA). Guinea-pig anti-insulin, rabbit anti-somatostatin (both from Dako, Glostrup, Denmark) were used at 1:100 dilution whilst mouse anti-glucagon (Sigma-Aldrich, St Louis, MO) was used at 1:400 dilution. Alexa-Fluor 488, 546, 633 F(ab')2 secondary antibodies (Thermo Fisher Scientific, Waltham, MA) were used at 1:200 dilution. Hoechst 33342, was used to visualize nuclei. Somatostatin was not assessed in brain samples. Immunostained cell preparations were scanned and assessed using a Zeiss LSM 510 laser confocal microscope (Zeiss, Oberkochen, Germany) after setting their thresholds below saturation with constant laser power and acquisition parameters across all samples.

Circos plot

The circos plot was generated using Circos v0.69-4 (17 Dec 2016) running on Perl 5.018002. The circos data visualization tool is based on the work by Krzywinski, M. et al.(Krzywinski et al., 2009) as well as on-line tutorials (http://circos.ca/). Results are presented in four different plots. Results are presented in four plots; (i) Scatter type plot illustrates the cycle threshold data. The plot is inverted so that the smaller values are at the top (further away from the center). (ii) Line plot presents z-scores of the expression data. Z-scores were calculated as the observed microRNA Ct values minus the average Ct-value for that microRNA divided by standard deviation of the samples and multiplied by -1. (iii) Histogram plot shows the relative z-scores (relative to the insulin-negative samples, n=39: comprising of 34 spleen, 2 muscle and one of each of the endothelial, liver, and skin donor tissues). The innermost plot displays correlations between miRNAs and mRNA levels. (iv) Pearson correlation coefficient is calculated for each of the miRNA – mRNA pairs with a cut-off set at Pearson's r=0.6. Correlations are plotted in the form of colored link (as per the labeled colors of the mRNA).

RNA-seq analysis

RNA-seq analysis was carried out using the Strand Next-generation sequencing (NGS) version 2.5 software on the human islet RNA-seq dataset (GSE134068, n=18) as described in (please refer to the accompanying related manuscript provided in confidence for review). Briefly, RNA-seq data were aligned to the human hg38 transcriptome and genome with unique splice variants through using Ensembl and transcript model and filtered based on base and read quality. DEseq was used to quantify and normalized the reads. Sequencing data are available from study accession GSE134068.

Statistical Analysis

Unsupervised bidirectional (cases-samples/variables-microRNAs) hierarchical clustering analysis involved average linkage and Euclidean distances between cases/variables to draw the heatmap using R package gplots (3.1.0) heatmap.2 function and RColorBrewer (1.1.2). Microsoft Excel ver. 2016 was used to generate the volcano plot by calculate Ct difference and p-values. Ct difference was obtained by subtracting average values for each microRNA between the two groups, while p-values were obtained using two-tailed t-test. Multiple testing using Benjamini-Hochberg approach to derive a critical value was performed in excel. L1-penalized logistic and linear regression techniques are used as described elsewhere(Goeman, 2010). This approach prevents overfitting of collinear and high-dimensional data and utilizes the LASSO (least absolute shrinkage and selection operator) algorithm. Penalized regression was used to derive a microRNA signature that is associated with (pro-)insulin expression. Model selection was performed using the LASSO method. Penalty applied to the regression coefficients allows for improving the predictive power and interpretability of regression models by selecting only a subset of all the available independent variables (microRNAs) rather than using all of them. Penalized regression analysis was carried out using logistic (dependent variable defined as presence (1) vs. absence (0) of (pro-)insulin expression) or linear (laboratory-measured Ct-value of (pro-)insulin gene and microRNA transcripts) regression workflow. The microRNAs identified by penalized regression analyses were validated using bootstrapping (n=1000 iterations) via the R package glmnet (4.0-2). Approximately 37% of samples are randomly eliminated and replaced with the same number of randomly selected samples within the set in each iteration for analysis. This proportion is the attribute of simple random sampling with replacement. When performing such sampling many times, on average ~63.2% of the observations would be obtained for each subset. This approach is based on the previous sampling methods (Efron and Tibshirani, 1997, Chernick and LaBudde, 2011) and

has been used in our previous study (Wong et al., 2019). Penalized regression analysis was generated using the R package penalized (0.9.51). ROC curve analysis was performed using R package pROC (1.16.2). T-Distributed Stochastic Neighbor Embedding (t-SNE) analysis and t-SNE plots creation were performed using R packages plyr (1.8.6), Rtsne (0.15), ggplot2 (3.3.2), devtools (2.3.2), tidyverse (1.3.0), colorspace (1.4.1), ggthemes (4.2.0) and scales (1.1.1). A perplexity of 30 with 1000 iterations were used for each t-SNE analysis. A categorical bubble plot was generated using the R packages ggplot2 (3.3.2), ggpubr (0.4.0) and proto (1.0.0). Correlation plot and analysis was generated using the R packages corrplot (0.84), Hmisc (4.4.1) and dplyr (1.0.2). All data presented herein were analyzed using Statistica for Windows ver. 13 (Dell Inc. Tulsa, OK), Microsoft Excel (Microsoft, Redmond, WA, USA), XLStat (Adinsoft, Paris, France), GraphPad Prism 8.4.1 (GraphPad Software, San Diego, CA, USA) and/or the R software ver. 3.6.2 (R Foundation for Statistical Computing, Vienna, Austria).

Pathway analysis

Mature miRNA sequences were obtained using miRBase. The validated and predicted gene targets for each miRNA were obtained using the computational web tool TargetScan Human 7.2 (http://www.targetscan.org/vert_72/)(Agarwal et al., 2015). KEGG Pathway analysis was performed in the miRSystem(Lu et al., 2012). The name of each selected microRNA was individually entered in both tools. In TargetScan, all predicted targets (irrespective of site conservation) are selected using the cumulative weighted context++ score as a guide to predict the interaction between the mRNA and microRNA. While in miRSystem, default parameters Hit ($\geq$3) and Default O/E ratio ($\geq$2) were used, with minimum total genes in pathways set at 4. KEGG pathways ranked with raw p-value converted to -log10(p-value), with a cut-off at $p \leq 0.05$ (or $\geq$1.30103 in -log10) are presented.

References

AGARWAL, V., BELL, G. W., NAM, J. W. & BARTEL, D. P. 2015. Predicting effective microRNA target sites in mammalian mRNAs. *Elife,* 4.

CHERNICK, M. R. & LABUDDE, R. 2011. *An Introduction to Bootstrap Methods with Applications,* New Jersey, Hoboken, John Wiley & Sons, Inc.

EFRON, B. & TIBSHIRANI, R. 1997. Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association,* 92**,** 548-560.

GERSHENGORN, M. C., HARDIKAR, A. A., WEI, C., GERAS-RAAKA, E., MARCUS-SAMUELS, B. & RAAKA, B. M. 2004. Epithelial-to-mesenchymal transition generates proliferative human islet precursor cells. *Science,* 306**,** 2261-4.

GOEMAN, J. J. 2010. L1 penalized estimation in the Cox proportional hazards model. *Biom J,* 52**,** 70-84.

HARDIKAR, A. A., FARR, R. J. & JOGLEKAR, M. V. 2014. Circulating microRNAs: understanding the limits for quantitative measurement by real-time PCR. *J Am Heart Assoc,* 3**,** e000792.

JOGLEKAR, M. V. & HARDIKAR, A. A. 2012. Isolation, expansion, and characterization of human islet-derived progenitor cells. *Methods Mol Biol,* 879**,** 351-66.

JOGLEKAR, M. V., JOGLEKAR, V. M., JOGLEKAR, S. V. & HARDIKAR, A. A. 2009. Human fetal pancreatic insulin-producing cells proliferate in vitro. *J Endocrinol,* 201**,** 27-36.

KRZYWINSKI, M., SCHEIN, J., BIROL, I., CONNORS, J., GASCOYNE, R., HORSMAN, D., JONES, S. J. & MARRA, M. A. 2009. Circos: an information aesthetic for comparative genomics. *Genome Res,* 19**,** 1639-45.

LU, T. P., LEE, C. Y., TSAI, M. H., CHIU, Y. C., HSIAO, C. K., LAI, L. C. & CHUANG, E. Y. 2012. miRSystem: an integrated system for characterizing enriched functions and pathways of microRNA targets. *PLoS One,* 7**,** e42390.

MAYNARD, C. L., WONG, W. K. M., HARDIKAR, A. A. & JOGLEKAR, M. V. 2019. Droplet Digital PCR for Measuring Absolute Copies of Gene Transcripts in Human Islet-Derived Progenitor Cells. *Methods Mol Biol,* 2029**,** 37-48.

MESTDAGH, P., VAN VLIERBERGHE, P., DE WEER, A., MUTH, D., WESTERMANN, F., SPELEMAN, F. & VANDESOMPELE, J. 2009. A novel and universal method for microRNA RT-qPCR data normalization. *Genome Biol,* 10**,** R64.

REN, B., O'BRIEN, B. A., SWAN, M. A., KOINA, M. E., NASSIF, N., WEI, M. Q. & SIMPSON, A. M. 2007. Long-term correction of diabetes in rats after lentiviral hepatic insulin gene therapy. *Diabetologia,* 50**,** 1910-20.

STUART, T., BUTLER, A., HOFFMAN, P., HAFEMEISTER, C., PAPALEXI, E., MAUCK, W. M., 3RD, HAO, Y., STOECKIUS, M., SMIBERT, P. & SATIJA, R. 2019. Comprehensive Integration of Single-Cell Data. *Cell,* 177**,** 1888-1902 e21.

WONG, W., FARR, R., JOGLEKAR, M., JANUSZEWSKI, A. & HARDIKAR, A. 2015. Probe-based Real-time PCR Approaches for Quantitative Measurement of microRNAs. *J Vis Exp*.

WONG, W. K., JIANG, G., SORENSEN, A. E., CHEW, Y. V., LEE-MAYNARD, C., LIUWANTARA, D., WILLIAMS, L., O'CONNELL, P. J., DALGAARD, L. T., MA, R. C., HAWTHORNE, W. J., JOGLEKAR, M. V. & HARDIKAR, A. A. 2019. The long noncoding RNA MALAT1 predicts human pancreatic islet isolation quality. *JCI Insight,* 5.